บทที่ 3
วิธีการดำเนินงานโครงงาน

โครงงานเรื่อง การวิเคราะห์ข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 เพื่อใช้สำหรับเผยแพร่ข้อมูลบนเว็บไซต์ ในบทนี้จะเป็นการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล (Data Mining) ซึ่งมีกระบวนการวิเคราะห์ข้อมูลด้วย CRISP–DM ที่สำคัญหลากหลายขั้นตอน เมื่อเสร็จสิ้นจากกระบวนการวิเคราะห์ข้อมูลแล้วจะเป็นการออกแบบเว็บไซต์ (Wireframe) และการออกแบบรูปแบบการแสดงผล (Dashboard) และบทสรุปจากวิธีการดำเนินงาน

3.1 การวิเคราะห์ข้อมูลด้วย CRISP–DM

3.2 การออกแบบเว็บไซต์

3.3 บทสรุป

3.1 การวิเคราะห์ข้อมูลด้วย CRISP–DM

กระบวนการวิเคราะห์ข้อมูล ด้วย CRISP–DM หรือ Cross Industry Standard Process for Data Mining พัฒนาขึ้นในปี ค.ศ. 1996 โดยความร่วมมือของ 3 บริษัทคือ Daimler Chrysler, SPSS และ NCR ประกอบด้วย 6 ขั้นตอนหลัก ได้แก่

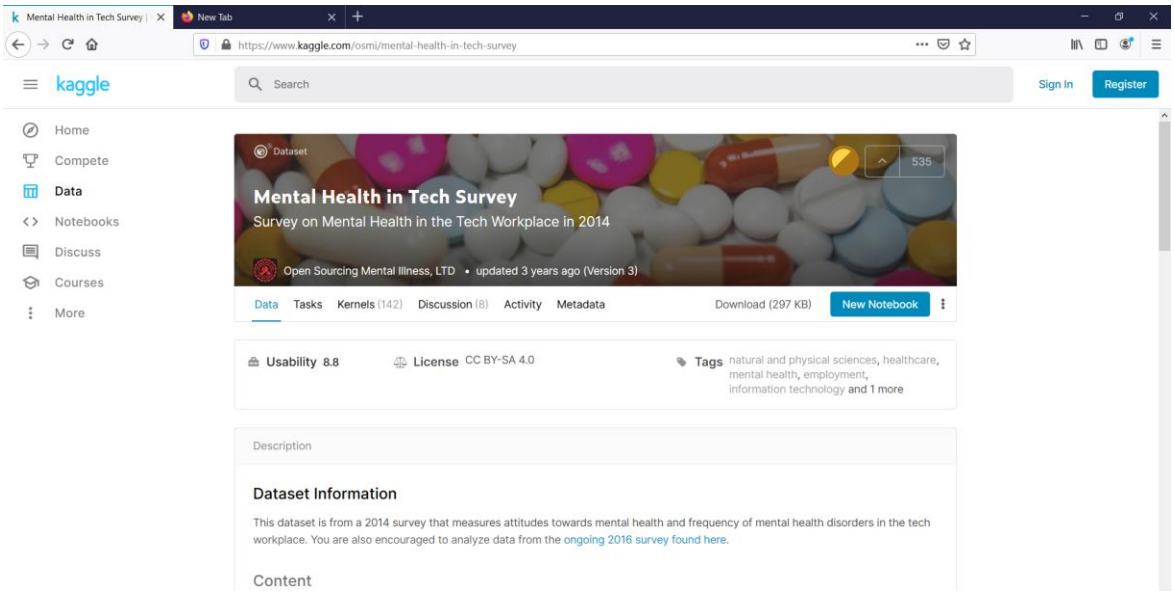
3.1.1 รู้จักและเข้าใจในธุรกิจ (Business Understanding) เป็นขั้นตอนแรกของกระบวนการที่มุ่งเน้นไปที่การทำความเข้าใจกระบวนการทางธุรกิจโดยรวม

ผู้วิเคราะห์ข้อมูลทำความเข้าใจกับปัญหาให้อยู่ในรูปของการวิเคราะห์ข้อมูลทาง Data Mining โดยการวิเคราะห์ข้อมูลในประเด็นนี้ คือ ข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 ซึ่งมีจำนวนข้อมูลมหาศาล จำนวนรายการทั้งหมด 1,259 รายการ ทำให้ไม่สามารถทำความเข้าใจกับข้อมูลที่มีความซับซ้อนได้อย่างรวดเร็ว เช่น ต้องการทราบว่าปัจจัยใดที่มีผลต่อปัญหาสุขภาพจิตในการสูงสุด

3.1.2 จัดเก็บและรวบรวมข้อมูลให้ครบ (Data Understanding) ขั้นตอนการจัดเก็บและรวบรวมข้อมูล ตลอดจนการพิจารณาตรวจสอบความถูกต้องของข้อมูลที่ได้รับ โดยเลือกว่าจะใช้ข้อมูลทั้งหมดหรือบางส่วนในการวิเคราะห์ให้สอดรับกับวัตถุประสงค์ที่กำหนดไว้

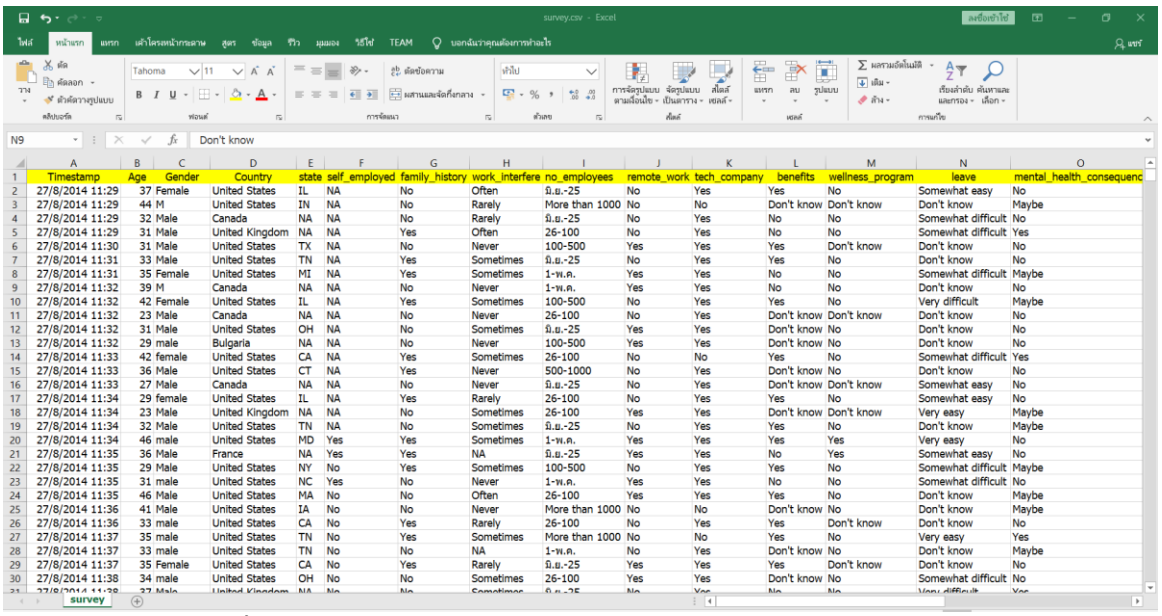
ผู้วิเคราะห์ข้อมูลทำการรวบรวมข้อมูล เพื่อตรวจสอบรายละเอียด ปริมาณ และความน่าเชื่อถือของข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 ที่ได้จาก

เว็บไซต์ www.kaggle.com ซึ่งเป็นระบบศูนย์กลางข้อมูลภาครัฐ ที่เผยแพร่สู่สาธารณะอย่างเป็นรูปธรรม เป็นช่องทางให้ผู้ใช้บริการทั้งภาคประชาชน ภาคธุรกิจเอกชน รวมถึงหน่วยงานของรัฐ สามารถค้นหา และเข้าถึงข้อมูลที่มีคุณภาพของภาครัฐได้ง่าย



ภาพที่ 3.1 เว็บไซต์ชุดข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559
(ที่มา : https://www.kaggle.com/osmi/mental-health-in-tech-survey)

ซึ่งข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 มีจำนวนข้อมูลทั้งหมด 1,259 รายการ ประกอบด้วย 19 แอตทริบิวท์ ประกอบด้วย เวลา (timestamp), อายุ (age), เพศ (gender), ประเทศ (country), รัฐ (state), เจ้าของกิจการ (self_employed), ครอบครัวเคยมีประวัติเกี่ยวกับปัญหาสุขภาพจิต (family_history), ปัญหาสุขภาพจิตรบกวนการทำงาน (work_interfere), จำนวนพนักงาน (no_employees), ทำงานในระยะไกล (remote_work), ทำงานเกี่ยวกับด้านเทคโนโลยี (tech_company), ได้สวัสดิการเกี่ยวกับสุขภาพจิต (benefits), มีโปรแกรมสุขภาพสำหรับพนักงาน (wellness_program), การลา (leave), สุขภาพจิตมีผลกระทบต่อสุขภาพ (mental health consequence), สุขภาพจิตมีผลกระทบต่อนายจ้าง (supervisor), สุขภาพจิตมีผลกระทบต่อเพื่อนร่วมงาน (coworkers), นายจ้างให้ความสำคัญกับสุขภาพจิตมากเท่ากับสุขภาพร่างกาย (mental vs physical) และผลการรักษา (treatment)



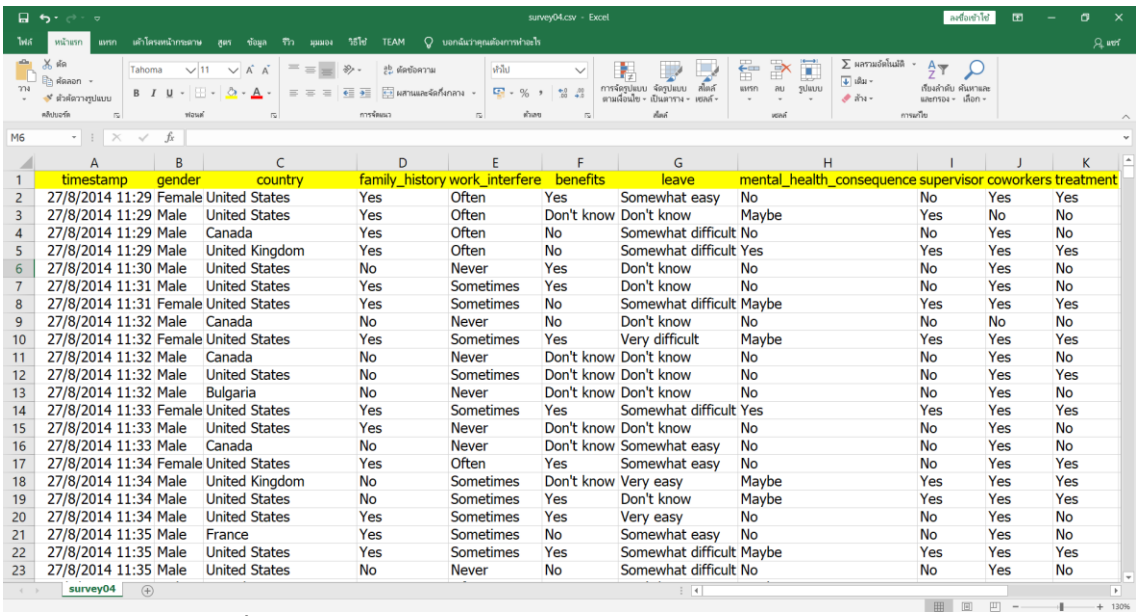
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Timestamp	Age	Gender	Country	state	self_employed	family_history	work_interfere	no_employees	remote_work	tech_company	benefits	wellness_program	leave	mental_health_consequenc
2	27/8/2014 11:29	37	Female	United States	IL	NA	No	Often	น.ย.-25	No	Yes	Yes	No	Somewhat easy	No
3	27/8/2014 11:29	44	M	United States	IN	NA	No	Rarely	More than 1000	No	No	Don't know	Don't know	Don't know	Maybe
4	27/8/2014 11:29	32	Male	Canada	NA	NA	No	Rarely	น.ย.-25	No	Yes	No	No	Somewhat difficult	No
5	27/8/2014 11:29	31	Male	United Kingdom	NA	NA	Yes	Often	26-100	No	Yes	No	No	Somewhat difficult	Yes
6	27/8/2014 11:30	31	Male	United States	TX	NA	No	Never	100-500	Yes	Yes	Yes	Don't know	Don't know	No
7	27/8/2014 11:31	33	Male	United States	TN	NA	Yes	Sometimes	น.ย.-25	No	Yes	Yes	No	Don't know	No
8	27/8/2014 11:31	35	Female	United States	MI	NA	Yes	Sometimes	1-พ.ค.	Yes	Yes	No	No	Somewhat difficult	Maybe
9	27/8/2014 11:32	39	M	Canada	NA	NA	No	Never	1-พ.ค.	Yes	Yes	No	No	Don't know	No
10	27/8/2014 11:32	42	Female	United States	IL	NA	Yes	Sometimes	100-500	No	Yes	Yes	No	Very difficult	Maybe
11	27/8/2014 11:32	23	Male	Canada	NA	NA	No	Never	26-100	No	Yes	Don't know	Don't know	Don't know	No
12	27/8/2014 11:32	31	Male	United States	OH	NA	No	Sometimes	น.ย.-25	Yes	Yes	Don't know	No	Don't know	No
13	27/8/2014 11:32	29	male	Bulgaria	NA	NA	No	Never	100-500	Yes	Yes	Don't know	No	Don't know	No
14	27/8/2014 11:33	42	female	United States	CA	NA	Yes	Sometimes	26-100	No	No	Yes	No	Somewhat difficult	Yes
15	27/8/2014 11:33	36	Male	United States	CT	NA	Yes	Never	500-1000	No	Yes	Don't know	No	Don't know	No
16	27/8/2014 11:33	27	Male	Canada	NA	NA	No	Never	น.ย.-25	No	Yes	Don't know	Don't know	Somewhat easy	No
17	27/8/2014 11:34	29	female	United States	IL	NA	Yes	Rarely	26-100	No	Yes	Yes	No	Somewhat easy	No
18	27/8/2014 11:34	23	Male	United Kingdom	NA	NA	No	Sometimes	26-100	Yes	Yes	Don't know	Don't know	Very easy	Maybe
19	27/8/2014 11:34	32	Male	United States	TN	NA	No	Sometimes	น.ย.-25	No	Yes	Yes	No	Don't know	Maybe
20	27/8/2014 11:34	46	male	United States	MD	Yes	Yes	Sometimes	1-พ.ค.	Yes	Yes	Yes	Yes	Very easy	No
21	27/8/2014 11:35	36	Male	France	NA	Yes	Yes	NA	น.ย.-25	Yes	Yes	No	Yes	Somewhat easy	No
22	27/8/2014 11:35	29	Male	United States	NY	No	Yes	Sometimes	100-500	No	Yes	Yes	No	Somewhat difficult	Maybe
23	27/8/2014 11:35	31	male	United States	NC	Yes	No	Never	1-พ.ค.	Yes	Yes	No	No	Somewhat difficult	No
24	27/8/2014 11:35	46	Male	United States	MA	No	No	Often	26-100	Yes	Yes	Yes	No	Don't know	Maybe
25	27/8/2014 11:36	41	Male	United States	IA	No	No	Never	More than 1000	No	No	Don't know	No	Don't know	Maybe
26	27/8/2014 11:36	33	male	United States	CA	No	Yes	Rarely	26-100	No	Yes	Yes	Don't know	Don't know	No
27	27/8/2014 11:37	35	male	United States	TN	No	Yes	Sometimes	More than 1000	No	No	Yes	No	Very easy	Yes
28	27/8/2014 11:37	33	male	United States	TN	No	No	NA	1-พ.ค.	No	Yes	Don't know	No	Don't know	Maybe
29	27/8/2014 11:37	35	Female	United States	CA	No	Yes	Rarely	น.ย.-25	Yes	Yes	Yes	Don't know	Don't know	No
30	27/8/2014 11:38	34	male	United States	OH	No	No	Sometimes	26-100	Yes	Yes	Don't know	No	Somewhat difficult	No

ภาพที่ 3.2 ข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559

3.1.3 เตรียมข้อมูลให้พร้อมใช้งาน (Data Preparation) ขั้นตอนการแปลงข้อมูลที่ได้รวบรวมมาและเลือกไว้ ให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปวิเคราะห์ในขั้นตอนต่อไปได้ โดยการทำให้เป็นข้อมูลที่ถูกต้อง (Data cleaning) มักใช้เวลาค่อนข้างมาก โดยมีขั้นตอนดังนี้

3.1.3.1 ทำการคัดเลือกข้อมูล (Data Selection) คือการคัดเลือกข้อมูลที่เหมาะสมเพื่อนำมาใช้ในการวิเคราะห์ข้อมูล

ผู้วิเคราะห์ข้อมูลทำการคัดเลือกข้อมูล และทำการ Data Cleaning ข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 โดยตัดส่วนที่ไม่จำเป็นออกให้เหลือเฉพาะข้อมูลที่จำเป็นในการวิเคราะห์ในภาพรวม จำนวน 11 แอตทริบิวท์ ได้แก่ เวลา (timestamp), เพศ (gender), ประเทศ (country), ครอบครัวเคยมีประวัติเกี่ยวกับปัญหาสุขภาพจิต (family_history), ปัญหาสุขภาพจิตรบกวนการทำงาน (work_interfere), ได้สวัสดิการเกี่ยวกับสุขภาพจิต (benefits), การลา (leave), สุขภาพจิตมีผลกระทบต่อสุขภาพ (mental health consequence), สุขภาพจิตมีผลกระทบต่อนายจ้าง (supervisor), สุขภาพจิตมีผลกระทบต่อเพื่อนร่วมงาน (coworkers), และผลการรักษา (treatment)



	timestamp	gender	country	family_history	work_interfere	benefits	leave	mental_health_consequence	supervisor	coworkers	treatment
2	27/8/2014 11:29	Female	United States	Yes	Often	Yes	Somewhat easy	No	No	Yes	Yes
3	27/8/2014 11:29	Male	United States	Yes	Often	Don't know	Don't know	Maybe	Yes	No	No
4	27/8/2014 11:29	Male	Canada	Yes	Often	No	Somewhat difficult	No	No	Yes	No
5	27/8/2014 11:29	Male	United Kingdom	Yes	Often	No	Somewhat difficult	Yes	Yes	Yes	Yes
6	27/8/2014 11:30	Male	United States	No	Never	Yes	Don't know	No	No	Yes	No
7	27/8/2014 11:31	Male	United States	Yes	Sometimes	Yes	Don't know	No	No	Yes	No
8	27/8/2014 11:31	Female	United States	Yes	Sometimes	No	Somewhat difficult	Maybe	Yes	Yes	Yes
9	27/8/2014 11:32	Male	Canada	No	Never	No	Don't know	No	No	No	No
10	27/8/2014 11:32	Female	United States	Yes	Sometimes	Yes	Very difficult	Maybe	Yes	Yes	Yes
11	27/8/2014 11:32	Male	Canada	No	Never	Don't know	Don't know	No	No	Yes	No
12	27/8/2014 11:32	Male	United States	No	Sometimes	Don't know	Don't know	No	No	Yes	Yes
13	27/8/2014 11:32	Male	Bulgaria	No	Never	Don't know	Don't know	No	No	Yes	No
14	27/8/2014 11:33	Female	United States	Yes	Sometimes	Yes	Somewhat difficult	Yes	Yes	Yes	Yes
15	27/8/2014 11:33	Male	United States	Yes	Never	Don't know	Don't know	No	No	Yes	No
16	27/8/2014 11:33	Male	Canada	No	Never	Don't know	Somewhat easy	No	No	Yes	No
17	27/8/2014 11:34	Female	United States	Yes	Often	Yes	Somewhat easy	No	No	Yes	Yes
18	27/8/2014 11:34	Male	United Kingdom	No	Sometimes	Don't know	Very easy	Maybe	Yes	Yes	Yes
19	27/8/2014 11:34	Male	United States	No	Sometimes	Yes	Don't know	Maybe	Yes	Yes	Yes
20	27/8/2014 11:34	Male	United States	Yes	Sometimes	Yes	Very easy	No	No	Yes	No
21	27/8/2014 11:35	Male	France	Yes	Sometimes	No	Somewhat easy	No	No	Yes	No
22	27/8/2014 11:35	Male	United States	Yes	Sometimes	Yes	Somewhat difficult	Maybe	Yes	Yes	Yes
23	27/8/2014 11:35	Male	United States	No	Never	No	Somewhat difficult	No	No	Yes	No

ภาพที่ 3.3 ข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559

3.1.3.2 ทำการกลั่นกรองข้อมูล (Data Cleaning) คือการทำความสะอาดข้อมูล เป็นกระบวนการตรวจสอบและการแก้ไข (หรือลบ) รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล ตารางหรือฐานข้อมูล ซึ่งเป็นหลักสำคัญของฐานข้อมูล ทางผู้วิเคราะห์ข้อมูลได้ดำเนินการดังนี้

1) ข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 ผู้วิเคราะห์ข้อมูลพบว่า ข้อมูลของเพศ (gender) ที่กรอกมามีความหลากหลาย เนื่องด้วยมาจากการคีย์ข้อมูลของต้นทางที่มีเพศสภาพแตกต่างกันในแต่ละบุคคล ดังนั้นผู้วิเคราะห์ข้อมูลจึงได้ทำการเพิ่มอีกคอลัมน์ขึ้นมา คือคอลัมน์ เพศ (sex) แล้วแก้ไขข้อมูลที่มีความหลากหลายจากต้นทางนั้น เพื่อจำแนกเพศจากข้อมูลตามหลักความเป็นจริงคือ เพศชาย และ เพศหญิง ดังนี้

- เพศหญิง ผู้วิเคราะห์ข้อมูลพบว่ามีการใช้ชื่อที่มีความหลากหลาย จึงได้ทำการตรวจสอบข้อมูลและทำการแก้ไขให้เป็นชื่อ Female ทั้งหมด

	timestamp	gender	sex	country	family_history
2	27/8/2014 11:29	Female	Female	United States	Yes
8	27/8/2014 11:31	Female	Female	United States	Yes
10	27/8/2014 11:32	Female	Female	United States	Yes
14	27/8/2014 11:33	female	Female	United States	Yes
17	27/8/2014 11:34	female	Female	United States	Yes
29	27/8/2014 11:37	Female	Female	United States	Yes
36	27/8/2014 11:40	female	Female	United States	No
47	27/8/2014 11:43	Female	Female	United States	Yes
63	27/8/2014 11:49	Female	Female	Poland	No
68	27/8/2014 11:51	Female	Female	Australia	Yes
70	27/8/2014 11:52	Female	Female	United States	No
74	27/8/2014 11:54	Trans-female	Female	United States	Yes
85	27/8/2014 12:03	Cis Female	Female	United States	Yes
89	27/8/2014 12:10	female	Female	United States	Yes
90	27/8/2014 12:11	F	Female	United States	No
96	27/8/2014 12:15	F	Female	Canada	No
102	27/8/2014 12:18	female	Female	Canada	Yes
103	27/8/2014 12:18	Female	Female	United States	No
140	27/8/2014 12:37	Female	Female	United States	No
146	27/8/2014 12:39	Female	Female	United States	Yes
151	27/8/2014 12:40	Female	Female	United States	No
163	27/8/2014 12:44	F	Female	United States	Yes
166	27/8/2014 12:46	F	Female	United States	Yes
167	27/8/2014 12:47	Woman	Female	United Kingdom	Yes
175	27/8/2014 12:49	Female	Female	United States	No
176	27/8/2014 12:49	Female	Female	Canada	Yes
179	27/8/2014 12:50	Female	Female	United States	Yes
180	27/8/2014 12:51	F	Female	United States	No

test_survey

ภาพที่ 3.4 ข้อมูลผิดพลาดจากการใช้ชื่อเพศหญิงที่มีความหลากหลาย

- เพศชาย ผู้วิเคราะห์ข้อมูลพบว่ามีการใช้ชื่อที่มีความหลากหลาย จึงได้ทำการตรวจสอบข้อมูลและทำการแก้ไขให้เป็นชื่อ Male ทั้งหมด

	timestamp	gender	sex	country	family_history
3	27/8/2014 11:29	M	Male	United States	Yes
4	27/8/2014 11:29	Male	Male	Canada	Yes
5	27/8/2014 11:29	Male	Male	United Kingdom	Yes
6	27/8/2014 11:30	Male	Male	United States	No
7	27/8/2014 11:31	Male	Male	United States	Yes
9	27/8/2014 11:32	M	Male	Canada	No
11	27/8/2014 11:32	Male	Male	Canada	No
12	27/8/2014 11:32	Male	Male	United States	No
13	27/8/2014 11:32	male	Male	Bulgaria	No
15	27/8/2014 11:33	Male	Male	United States	Yes
16	27/8/2014 11:33	Male	Male	Canada	No
18	27/8/2014 11:34	Male	Male	United Kingdom	No
19	27/8/2014 11:34	Male	Male	United States	No
20	27/8/2014 11:34	male	Male	United States	Yes
21	27/8/2014 11:35	Male	Male	France	Yes
22	27/8/2014 11:35	Male	Male	United States	Yes
23	27/8/2014 11:35	male	Male	United States	No
24	27/8/2014 11:35	Male	Male	United States	Yes
25	27/8/2014 11:36	Male	Male	United States	No
26	27/8/2014 11:36	male	Male	United States	Yes
27	27/8/2014 11:37	male	Male	United States	Yes
28	27/8/2014 11:37	male	Male	United States	No
30	27/8/2014 11:38	male	Male	United States	No
31	27/8/2014 11:38	Male	Male	United Kingdom	No
32	27/8/2014 11:39	Male	Male	United Kingdom	No
33	27/8/2014 11:38	Male	Male	United States	Yes
34	27/8/2014 11:39	male	Male	United Kingdom	Yes
35	27/8/2014 11:39	Male	Male	United States	Yes

test_survey

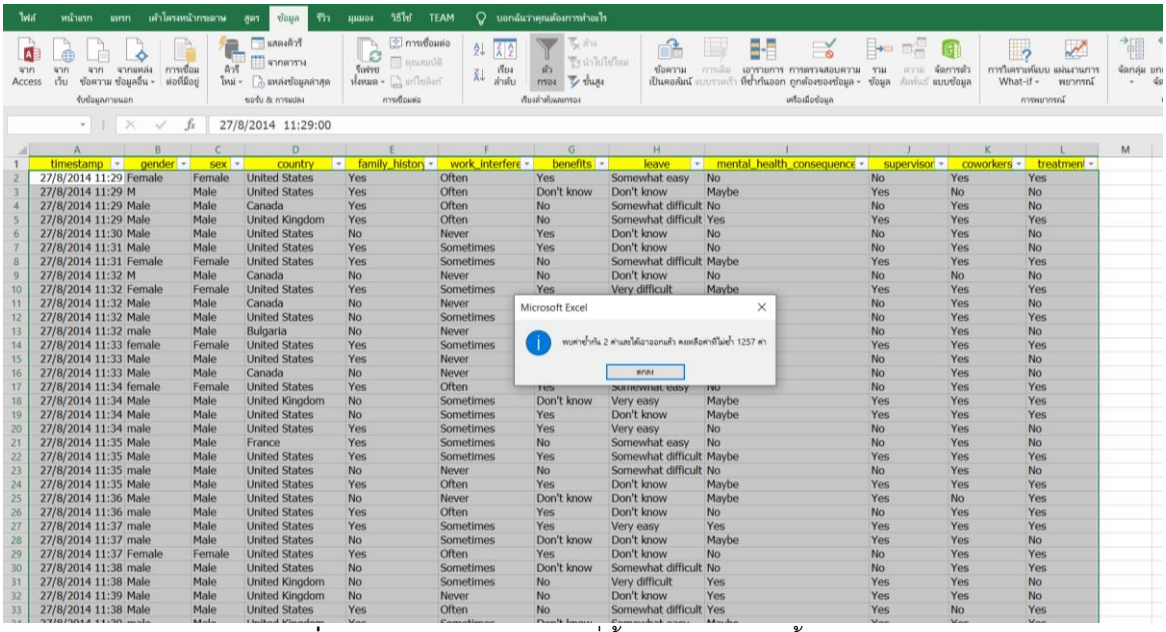
ภาพที่ 3.5 ข้อมูลผิดพลาดจากการใช้ชื่อเพศชายที่มีความหลากหลาย

- สำหรับเพศอื่น ๆ ผู้วิเคราะห์ข้อมูลพบว่ามีการใช้ชื่อเพศอื่นที่ไม่สามารถระบุเพศได้ จึงได้ทำการตรวจสอบข้อมูลและทำการแก้ไขให้เป็นชื่อ Other ทั้งหมด

1	timestamp	gender	sex	country	family_history
95	27/8/2014 12:15	something kinda ma	Other	Russia	No
308	27/8/2014 14:22	queer/she/they	Other	United States	Yes
351	27/8/2014 14:47	non-binary	Other	United States	Yes
389	27/8/2014 15:24	Nah	Other	United States	Yes
392	27/8/2014 15:24	All	Other	Zimbabwe	Yes
415	27/8/2014 15:30	Enby	Other	United Kingdom	Yes
418	27/8/2014 15:31	fluid	Other	Germany	Yes
453	27/8/2014 15:53	Genderqueer	Other	United States	Yes
546	27/8/2014 18:13	Androgyne	Other	United Kingdom	Yes
555	27/8/2014 18:56	Agender	Other	United Kingdom	No
779	28/8/2014 12:10	Neuter	Other	United Kingdom	No
820	28/8/2014 14:46	queer	Other	Germany	No
1259					

ภาพที่ 3.6 ข้อมูลผิดพลาดจากการใช้ชื่อเพศอื่นที่ไม่สามารถระบุเพศได้

ผู้วิเคราะห์ข้อมูลทำการตรวจสอบข้อมูลที่ซ้ำซ้อนของข้อมูลทั้งหมด พบข้อมูลที่ซ้ำกันจำนวน 2 รายการ ทางผู้วิเคราะห์ข้อมูลจึงได้ลบข้อมูลนั้นทิ้ง โดยจะมีข้อมูลในปี 2557–2559 ที่ทำการ Cleaning แล้ว จำนวน 1,257 รายการ



ภาพที่ 3.7 การตรวจสอบข้อมูลที่ซ้ำซ้อนของข้อมูลทั้งหมด

3.1.3.3 แปลงรูปแบบของข้อมูล (Data Transformation) เป็นขั้นตอนการแปลงข้อมูลในรูปแบบตารางฐานข้อมูลให้อยู่ในรูป item set เพื่อใช้สำหรับการวิเคราะห์ด้วยวิธีการของ data mining ทางผู้วิเคราะห์ข้อมูลได้ดำเนินการกับข้อมูลปัญหาสุขภาพจิตในการทำงานประจำปี 2557 – 2559 ดังนี้

1) ผู้วิเคราะห์ข้อมูลพบข้อมูล “ไม่ทราบ (Don't know)” ในคอลัมน์ของแต่ละประเภท คือ ได้สวัสดิการเกี่ยวกับสุขภาพจิต (benefits), มีโปรแกรมสุขภาพสำหรับพนักงาน (wellness_program) และการลา (leave) จึงได้ทำการแปลงข้อความให้อยู่ในรูปแบบตัวเลขเพื่อหาฐานนิยม หรือตัวเลขที่ซ้ำกันมากที่สุดในช่วงของข้อมูล ซึ่งจะได้ตรงกับข้อความที่มีการเกิดขึ้นมากที่สุด ผู้วิเคราะห์ข้อมูลจึงนำมาแทนที่ประเภทไม่ทราบของแต่ละกลุ่มข้อมูลทั้งหมด ด้วยฟังก์ชัน MODE

2) ผู้วิเคราะห์ข้อมูลทำการแปลงรูปแบบด้วยรวมกลุ่มของข้อมูลในแอตทริบิวต์ สุขภาพจิตรบกวนการทำงาน (work_interfere) เนื่องจากปกติแล้วจะมีข้อมูลในแอตทริบิวต์นี้หลากหลาย ได้แก่ Often, NA, Never, Rarely, Sometimes ซึ่งผู้วิจัยจะขอรวมกลุ่มของสุขภาพจิตรบกวนการทำงาน (work_interfere) ในรูปแบบ NA และ Never ให้เป็น Never และในรูปแบบ Rarely และ Sometimes ให้เป็น Sometimes เพียงอย่างเดียว

3.1.4 สร้างแบบจำลอง (Modeling) ขั้นตอนการสร้างตัวแบบทางคณิตศาสตร์ และสถิติเพื่อการวิเคราะห์ข้อมูล โดยสามารถใช้เทคนิควิธีการต่าง ๆ อาทิ การจำแนก (Classification) การแบ่งกลุ่ม (Clustering) และการสร้างความสัมพันธ์ (Association rule)

ผู้วิเคราะห์ข้อมูล ได้วิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูลแบบ Data Classification เพื่อใช้ทำนายแนวโน้มการเกิดขึ้นของปัจจัยที่ก่อให้เกิดปัญหาสุขภาพจิตในการทำงานจากเทคนิคการจำแนกประเภทข้อมูล ด้วยการสร้างโมเดล Decision Tree เพื่อจัดกลุ่มข้อมูลที่มีความสัมพันธ์กัน โดยใช้โปรแกรมที่ใช้ทำเหมืองข้อมูล ด้วยชุดข้อมูลที่คัดเลือก ดังนี้

family_history	work_interfere	supervisor	coworkers	treatment
Yes	Often	No	Yes	Yes
Yes	Often	Yes	No	No
Yes	Often	No	Yes	No
Yes	Often	Yes	Yes	Yes
No	Never	No	Yes	No
Yes	Sometimes	No	Yes	No
Yes	Sometimes	Yes	Yes	Yes
No	Never	No	No	No
Yes	Sometimes	Yes	Yes	Yes
No	Never	No	Yes	No
No	Sometimes	No	Yes	Yes
No	Never	No	Yes	No
Yes	Sometimes	Yes	Yes	Yes
Yes	Never	No	Yes	No
No	Never	No	Yes	No
Yes	Often	No	Yes	Yes
No	Sometimes	Yes	Yes	Yes
No	Sometimes	Yes	Yes	Yes
Yes	Sometimes	No	Yes	No
Yes	Sometimes	No	Yes	No
Yes	Sometimes	Yes	Yes	Yes
No	Never	No	Yes	No
Yes	Often	Yes	Yes	Yes
No	Never	Yes	No	Yes
Yes	Often	No	Yes	Yes

ภาพที่ 3.8 ชุดข้อมูลที่คัดเลือกมาวิเคราะห์ข้อมูล

จากรูปภาพที่ ประกอบด้วย 5 แอตทริบิวต์ คือ

- ประวัติครอบครัว (family history) แสดงประวัติครอบครัวที่เคยมีปัญหาสุขภาพจิต

ประกอบด้วย 2 ค่า คือ ใช่ ไม่ใช่ (Yes/No)

- การรบกวนการทำงาน (work interfere) แสดงสุขภาพจิตในการรบกวนการทำงาน ประกอบด้วย 3 ค่า คือ บ่อยครั้ง บางครั้ง ไม่เคย (Often/Sometimes/Never)
- มีผลกระทบต่อนายจ้าง (supervisor) แสดงสุขภาพจิตมีผลกระทบต่อนายจ้าง ประกอบด้วย 2 ค่า คือ ใช่ ไม่ใช่ (Yes/No)
- มีผลกระทบต่อเพื่อนร่วมงาน (coworkers) แสดงสุขภาพจิตมีผลกระทบต่อเพื่อน

ร่วมงาน ประกอบด้วย 2 ค่า คือ ใช่ ไม่ใช่ (Yes/No)

- ผลการรักษา (treatment) แสดงผลการรักษา ซึ่งเป็นคลาส ประกอบด้วย 2 ค่า คือ ควรไปรักษา และไม่ควรไปรักษา (Yes/No)

การสร้างโมเดล decision tree จะทำการคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของ tree (root node) หลังจากนั้นก็จะหาแอตทริบิวต์ถัดไปเรื่อย ๆ ในการหาความสัมพันธ์ของแอตทริบิวต์นี้จะใช้ตัววัด ที่เรียกว่า Information Gain (IG) ค่านี้คำนวณได้จากสมการดังนี้

ตารางที่ 3.1 สมการการหาความสัมพันธ์ของแอตทริบิวต์

การคำนวณ	**สมการ**
Entropy	Entropy(c1) = –p(c1) log p(c1) และ p(c2) คือ ค่าความน่าจะเป็นของ c1
Information Gain	IG (parent, child) = entropy(parent) – [p(c1) × entropy(c1) + p(c2) × entropy(c2) + ...]

การคำนวณค่าแต่ละแอตทริบิวต์เทียบกับคลาสเพื่อหาแอตทริบิวต์ที่มีค่า IG มากที่สุดมาเป็น root ของ decision tree กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นควรไปรักษา และไม่ควรไปรักษา ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ family history จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

$$\text{entropy (parent)} = -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$$

= −[0.52 × $\log_2$(0.52) + 0.48 × $\log_2$(0.48)]

= −[0.52 × −0.94 + 0.48 × −1.06]

= −[0.49 + 0.51]

= 1.00

entropy(ผล = Yes) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.74 × $\log_2$(0.74) + 0.26 × $\log_2$(0.26)]

= −[0.74 × −0.43 + 0.26 × −1.94]

= −[0.32 + 0.50]

= 0.82

entropy(ผล = No) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.28 × $\log_2$(0.28) + 0.72 × $\log_2$(0.72)]

= −[0.28 × −1.84 + 0.72 × −0.47]

= −[0.52 + 0.34]

= 0.86

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)+ p(ผล = No) × entropy(ผล = No)]

= 1.00 – [0.51 × 0.82 + 0.49 × 0.86]

= 1.00 – [0.42 + 0.42]

= 1.00 – 0.84

= 0.16

2) คำนวณค่า IG ของแอตทริบิวต์ work interfere จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.52 × $\log_2$(0.52) + 0.48 × $\log_2$(0.48)]

= −[0.52 × −0.94 + 0.48 × −1.06]

= −[0.49 + 0.51]

= 1.00

entropy(ผล = Often) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.80 × $\log_2$(0.80) + 0.20 × $\log_2$(0.20)]

$= -[0.80 \times -0.32 + 0.20 \times -2.32]$

$= -[0.26 + 0.46]$

$= 0.72$

entropy(ผล = Never) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.16 \times \log_2(0.16) + 0.84 \times \log_2(0.84)]$

$= -[0.16 \times -2.64 + 0.84 \times -0.25]$

$= -[0.42 + 0.21]$

$= 0.63$

entropy(ผล = Sometimes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.50 \times \log_2(0.50) + 0.50 \times \log_2(0.50)]$

$= -[0.50 \times -1.00 + 0.50 \times -1.00]$

$= -[0.50 + 0.50]$

$= 1.00$

IG(parent, child) = entropy (parent) – [p(ผล = Often) × entropy(ผล = Often) + p(ผล = Never) × entropy(ผล = Never) + p(ผล = Sometimes) × entropy(ผล = Sometimes)]

$= 1.00 - [0.25 \times 0.72 + 0.17 \times 0.63 + 0.58 \times 1.00]$

$= 1.00 - [0.18 + 0.11 + 0.58]$

$= 1.00 - 0.87$

$= 0.13$

3) คำนวณค่า IG ของแอตทริบิวต์ supervisor จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.52 \times \log_2(0.52) + 0.48 \times \log_2(0.48)]$

$= -[0.52 \times -0.94 + 0.48 \times -1.06]$

$= -[0.49 + 0.51]$

$= 1.00$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.57 \times \log_2(0.57) + 0.43 \times \log_2(0.43)]$

= −[0.57 × −0.81 + 0.43 × −1.22]

= −[0.46 + 0.52]

= 0.98

entropy(ผล =No) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.43 × $\log_2$(0.43) + 0.57 × $\log_2$(0.57)]

= −[0.43 × −1.22 + 0.57 × −0.82]

= −[0.52 + 0.47]

= 0.99

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes) + p(ผล = No) × entropy(ผล = No)]

= 1.00 – [0.61 × 0.98 + 0.39 × 0.99]

= 1.00 – [0.60 + 0.39]

= 1.00 – 0.99

= 0.01

4) คำนวณค่า IG ของแอตทริบิวต์ coworkers จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.52 × $\log_2$(0.52) + 0.48 × $\log_2$(0.48)]

= −[0.52 × −0.94 + 0.48 × −1.06]

= −[0.49 + 0.51]

= 1.00

entropy(ผล = Yes) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.53 × $\log_2$(0.53) + 0.47 × $\log_2$(0.47)]

= −[0.53 × −0.92 + 0.47 × −1.09]

= −[0.49 + 0.51]

= 1.00

entropy(ผล =No) = −p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= −[0.47 × $\log_2$(0.47) + 0.53 × $\log_2$(0.53)]

= −[0.47 × −1.09 + 0.53 × −0.92]

$= -[0.51 + 0.49]$

$= 1.00$

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes) + p(ผล = No) × entropy(ผล = No)]

$= 1.00 - [0.79 \times 1.00 + 0.21 \times 1.00]$

$= 1.00 - [0.79 + 0.21]$

$= 1.00 - 1.00$

$= 0.00$

จากการคำนวณค่า IG ของทุกแอตทริบิวต์พบว่าค่า IG ของแอตทริบิวต์ family history มีค่ามากที่สุด (0.16) ดังนั้นจึงเลือกแอตทริบิวต์ family history ขึ้นมาเป็นโหนด root และจะต้องทำการแตกกิ่งจากโหนด root ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกัน และผู้วิเคราะห์ข้อมูลพบว่าการคำนวณแอตทริบิวต์ family history(No) ไม่สามารถสร้างกิ่งแต่ละโหนดต่อไปได้ เนื่องจากไม่มีความสัมพันธ์กับแอตทริบิวต์ใด จึงสรุปข้อมูลได้เป็นผลลัพธ์ควรไปรักษา และไม่ควรไปรักษา ดังนั้นผู้วิเคราะห์ข้อมูลจึงสร้างโหนดในระดับถัดไปของแอตทริบิวต์ family history(yes)

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 2 ต่อจากโหนด root เพื่อหาค่า IG ที่มากที่สุด ของแอตทริบิวต์ family history กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นควรไปรักษา และไม่ควรไปรักษา ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ family history และแอตทริบิวต์ work interfere ของ Yes จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.74 \times \log_2(0.74) + 0.26 \times \log_2(0.26)]$

$= -[0.74 \times -0.43 + 0.26 \times -1.94]$

$= -[0.32 + 0.50]$

$= 0.82$

entropy(ผล = Often) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.80 \times \log_2(0.80) + 0.20 \times \log_2(0.20)]$

$= -[0.80 \times -0.32 + 0.20 \times -2.32]$

$= -[0.26 + 0.46]$

$= 0.72$

entropy(ผล = Never) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.39 \times \log_2(0.39) + 0.61 \times \log_2(0.61)]$

$= -[0.39 \times -1.36 + 0.61 \times -0.71]$

$= -[0.53 + 0.43]$

$= 0.96$

entropy(ผล = Sometimes) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.72 \times \log_2(0.72) + 0.28 \times \log_2(0.28)]$

$= -[0.72 \times -0.47 + 0.28 \times -1.84]$

$= -[0.34 + 0.52]$

$= 0.86$

IG(parent, child) = entropy (parent) – [p(ผล = Often) × entropy(ผล = Often) + p(ผล = Never) × entropy(ผล = Never) + p(ผล = Sometimes) × entropy (ผล = Sometimes)]

$= 0.82 - [0.49 \times 0.72 + 0.06 \times 0.96 + 0.45 \times 0.86]$

$= 0.82 - [0.35 + 0.06 + 0.39]$

$= 0.82 - 0.80$

$= 0.02$

1.1) คำนวณค่า IG ของแอตทริบิวต์ family history และแอตทริบิวต์ work interfere ของ No จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.28 \times \log_2(0.28) + 0.72 \times \log_2(0.72)]$

$= -[0.28 \times -1.84 + 0.72 \times -0.47]$

$= -[0.52 + 0.34]$

$= 0.86$

entropy(ผล = Often) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.00 \times \log_2(0.00) + 0.00 \times \log_2(0.00)]$

$= -[0.00 \times -0.00 + 0.00 \times -0.00]$

$= -[0.00 + 0.00]$

$= 0.00$

entropy(ผล = Never) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.11 \times \log_2(0.11) + 0.89 \times \log_2(0.89)]$

$= -[0.11 \times -3.18 + 0.89 \times -0.17]$

$= -[0.35 + 0.15]$

$= 0.50$

entropy(ผล = Sometimes) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.35 \times \log_2(0.35) + 0.65 \times \log_2(0.65)]$

$= -[0.35 \times -1.51 + 0.65 \times -0.62]$

$= -[0.53 + 0.40]$

$= 0.93$

IG(parent, child) = entropy (parent) – [p(ผล = Never) × entropy(ผล = Never) + p(ผล = Sometimes) × entropy (ผล = Sometimes)]

$= 0.86 - [0.00 \times 0.00 + 0.29 \times 0.50 + 0.71 \times 0.93]$

$= 0.86 - [0.00 + 0.15 + 0.66]$

$= 0.86 - 0.81$

$= 0.05$

2) คำนวณค่า IG ของแอตทริบิวต์ family history และแอตทริบิวต์ supervisor ของ Yes จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.74 \times \log_2(0.74) + 0.26 \times \log_2(0.26)]$

$= -[0.74 \times -0.43 + 0.26 \times -1.94]$

$= -[0.32 + 0.50]$

$= 0.82$

entropy(ผล = Yes) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

$= -[0.76 \times \log_2(0.76) + 0.24 \times \log_2(0.24)]$

$= -[0.76 \times -0.40 + 0.24 \times -2.06]$

$= -[0.30 + 0.49]$

= 0.79

entropy(ผล = No) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.69 \times \log_2(0.69) + 0.31 \times \log_2(0.31)]$

= $-[0.69 \times -0.54 + 0.31 \times -1.70]$

= $-[0.37 + 0.53]$

= 0.90

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)
+ p(ผล = No) × entropy(ผล = No)]

= 0.82 – [0.65 × 0.79 + 0.35 × 0.90]

= 0.82 – [0.51 + 0.32]

= 0.82 – 0.83

= –0.01

2.1) คำนวณค่า IG ของแอตทริบิวต์ family history และแอตทริบิวต์ supervisor ของ No จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.28 \times \log_2(0.28) + 0.72 \times \log_2(0.72)]$

= $-[0.28 \times -1.84 + 0.72 \times -0.47]$

= $-[0.52 + 0.34]$

= 0.86

entropy(ผล = Yes) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.33 \times \log_2(0.33) + 0.67 \times \log_2(0.67)]$

= $-[0.33 \times -1.60 + 0.67 \times -0.58]$

= $-[0.53 + 0.40]$

= 0.93

entropy(ผล = No) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.21 \times \log_2(0.21) + 0.79 \times \log_2(0.79)]$

= $-[0.21 \times -2.25 + 0.79 \times -0.34]$

= $-[0.47 + 0.27]$

= 0.74

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)

+ p(ผล = No) × entropy(ผล = No)]

= 0.86 – [0.57 × 0.93 + 0.43 × 0.74]

= 0.86 – [0.53 + 0.32]

= 0.86 – 0.85

= 0.01

3) คำนวณค่า IG ของแอตทริบิวต์ family history และแอตทริบิวต์ coworkers ของ Yes จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

= $-[0.74 \times \log_2(0.74) + 0.26 \times \log_2(0.26)]$

= $-[0.74 \times -0.43 + 0.26 \times -1.94]$

= –[0.32 + 0.50]

= 0.82

entropy(ผล = Yes) = $-p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

= $-[0.76 \times \log_2(0.76) + 0.24 \times \log_2(0.24)]$

= $-[0.76 \times -0.40 + 0.24 \times -2.06]$

= –[0.30 + 0.49]

= 0.79

entropy(ผล = No) = $-p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

= $-[0.64 \times \log_2(0.64) + 0.36 \times \log_2(0.36)]$

= $-[0.64 \times -0.64 + 0.36 \times -1.47]$

= –[0.41 + 0.53]

= 0.94

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)

+ p(ผล = No) × entropy(ผล = No)]

= 0.82 – [0.79 × 0.79 + 0.21 × 0.94]

= 0.82 – [0.62 + 0.20]

= 0.82 – 0.82

= 0.00

3.1) คำนวณค่า IG ของแอตทริบิวต์ family history และแอตทริบิวต์ coworkers ของ No จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.28 \times \log_2(0.28) + 0.72 \times \log_2(0.72)]$

$= -[0.28 \times -1.84 + 0.72 \times -0.47]$

$= -[0.52 + 0.34]$

$= 0.86$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.28 \times \log_2(0.28) + 0.72 \times \log_2(0.72)]$

$= -[0.28 \times -1.84 + 0.72 \times -0.47]$

$= -[0.52 + 0.34]$

$= 0.86$

entropy(ผล = No) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.27 \times \log_2(0.27) + 0.73 \times \log_2(0.73)]$

$= -[0.27 \times -1.89 + 0.73 \times -0.45]$

$= -[0.51 + 0.33]$

$= 0.84$

IG(parent, child) $=$ entropy (parent) $-$ [p(ผล = Yes) $\times$ entropy(ผล = Yes) + p(ผล = No) $\times$ entropy(ผล = No)]

$= 0.86 - [0.80 \times 0.86 + 0.20 \times 0.84]$

$= 0.86 - [0.69 + 0.17]$

$= 0.86 - 0.86$

$= 0.00$

จากการคำนวณค่า IG ของแอตทริบิวต์ family history(Yes) ต่อแอตทริบิวต์ work interfere แอตทริบิวต์ supervisor และแอตทริบิวต์ coworkers พบว่าค่า IG ของแอตทริบิวต์ family history ต่อแอตทริบิวต์ work interfere มีค่ามากที่สุด (0.05) ดังนั้นจึงเลือกแอตทริบิวต์ work interfere ขึ้นมาเป็นโหนดในระดับที่ 2 ต่อจากโหนด Root และทำการแตกกิ่งจากโหนดในระดับที่ 2 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกัน

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 3 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นควรไปรักษา และไม่ควรไปรักษา ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) และแอตทริบิวต์ work interfere(Never) ไปแอตทริบิวต์ supervisor(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.39 \times \log_2(0.39) + 0.61 \times \log_2(0.61)]$

$= -[0.39 \times -1.36 + 0.61 \times -0.71]$

$= -[0.53 + 0.43]$

$= 0.96$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.44 \times \log_2(0.44) + 0.56 \times \log_2(0.56)]$

$= -[0.44 \times -1.18 + 0.56 \times -0.84]$

$= -[0.52 + 0.47]$

$= 0.99$

entropy(ผล = No) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.27 \times \log_2(0.27) + 0.73 \times \log_2(0.73)]$

$= -[0.27 \times -1.89 + 0.73 \times -0.45]$

$= -[0.51 + 0.33]$

$= 0.84$

IG(parent, child) $=$ entropy (parent) $-$ [p(ผล = Yes) $\times$ entropy(ผล = Yes) + p(ผล = No) $\times$ entropy(ผล = No)]

$= 0.96 - [0.71 \times 0.99 + 0.29 \times 0.84]$

$= 0.96 - [0.70 + 0.24]$

$= 0.96 - 0.94$

$= 0.02$

2) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) และแอตทริบิวต์ work interfere(Never) ไปแอตทริบิวต์ coworkers(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.39 \times \log_2(0.39) + 0.61 \times \log_2(0.61)]$

$= -[0.39 \times -1.36 + 0.61 \times -0.71]$

$= -[0.53 + 0.43]$

$= 0.96$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.43 \times \log_2(0.43) + 0.57 \times \log_2(0.57)]$

$= -[0.43 \times -1.22 + 0.57 \times -0.81]$

$= -[0.52 + 0.46]$

$= 0.98$

entropy(ผล = No) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.33 \times \log_2(0.33) + 0.67 \times \log_2(0.67)]$

$= -[0.33 \times -1.60 + 0.67 \times -0.58]$

$= -[0.53 + 0.39]$

$= 0.92$

IG(parent, child) $=$ entropy (parent) $-$ [p(ผล = Yes) $\times$ entropy(ผล = Yes) + p(ผล = No) $\times$ entropy(ผล = No)]

$= 0.96 - [0.61 \times 0.98 + 0.39 \times 0.92]$

$= 0.96 - [0.60 + 0.36]$

$= 0.96 - 0.96$

$= 0.00$

จากการคำนวณค่า IG ของแอตทริบิวต์ family history(Yes) ต่อแอตทริบิวต์ work interfere(Never) ไปแอตทริบิวต์ supervisor และแอตทริบิวต์ coworkers พบว่าค่า IG ของแอตทริบิวต์ family history ต่อแอตทริบิวต์ work interfere ไปแอตทริบิวต์ supervisor มีค่ามากที่สุด (0.02) ดังนั้นจึงเลือกแอตทริบิวต์ supervisor ขึ้นมาเป็นโหนดในระดับที่ 3 ต่อจากโหนด Root ต่อจากโหนดระดับที่ 2 และทำการแตกกิ่งจากโหนดในระดับที่ 3 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกัน

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 4 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นควรไปรักษา และไม่ควรไปรักษา ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) แอตทริบิวต์ work interfere(Naver) และแอตทริบิวต์ supervisor(No) ไปแอตทริบิวต์ cowoker(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.27 \times \log_2(0.27) + 0.73 \times \log_2(0.73)]$

$= -[0.27 \times -1.89 + 0.73 \times -0.45]$

$= -[0.51 + 0.33]$

$= 0.84$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.20 \times \log_2(0.20) + 0.80 \times \log_2(0.80)]$

$= -[0.20 \times -2.32 + 0.80 \times -0.32]$

$= -[0.46 + 0.26]$

$= 0.72$

entropy(ผล = No) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[1.00 \times \log_2(1.00) + 0.00 \times \log_2(0.00)]$

$= -[1.00 \times -0 + 0.00 \times -0.00]$

$= -[0.00 + 0.00]$

$= 0.00$

IG(parent, child) $=$ entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes) + p(ผล = No) × entropy(ผล = No)]

$= 0.84 - [0.91 \times 0.72 + 0.09 \times 0.00]$

$= 0.84 - [0.66 + 0.00]$

$= 0.84 - 0.66$

$= 0.18$

2) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) แอตทริบิวต์ work interfere(Naver) และแอตทริบิวต์ supervisor(Yes) ไปแอตทริบิวต์ cowoker(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.44 \times \log_2(0.44) + 0.56 \times \log_2(0.56)]$

$= -[0.44 \times -1.18 + 0.73 \times -0.84]$

$= -[0.52 + 0.61]$

$= 1.13$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.62 \times \log_2(0.62) + 0.38 \times \log_2(0.38)]$

$= -[0.62 \times -0.69 + 0.38 \times -1.40]$

$= -[0.43 + 0.55]$

$= 0.98$

entropy(ผล = No) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.29 \times \log_2(0.29) + 0.71 \times \log_2(0.71)]$

$= -[0.29 \times -1.79 + 0.71 \times -0.49]$

$= -[0.52 + 0.35]$

$= 0.87$

IG(parent, child) $=$ entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)
+ p(ผล = No) × entropy(ผล = No)]

$= 1.13 - [0.48 \times 0.98 + 0.52 \times 0.87]$

$= 1.13 - [0.47 + 0.45]$

$= 1.13 - 0.92$

$= 0.21$

จากการคำนวณค่า IG ของแอตทริบิวต์ family history(Yes) แอตทริบิวต์ work interfere(Naver) แอตทริบิวต์ supervisor(Yes) แอตทริบิวต์ cowoker(Yes และ No) พบว่าข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันแล้ว คือ ผลลัพธ์เป็นควรไปรักษาและไม่ควรไปรักษา

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 3 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นควรไปรักษา และไม่ควรไปรักษา ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) และแอตทริบิวต์ work interfere (Sometimes) ไปแอตทริบิวต์ supervisor(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.72 \times \log_2(0.72) + 0.28 \times \log_2(0.28)]$

$= -[0.72 \times -0.47 + 0.28 \times -1.84]$

$= -[0.34 + 0.52]$

= 0.86

entropy(ผล = Yes) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.77 \times \log_2(0.77) + 0.23 \times \log_2(0.23)]$

= $-[0.77 \times -0.38 + 0.23 \times -2.12]$

= $-[0.29 + 0.49]$

= 0.78

entropy(ผล = No) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.62 \times \log_2(0.62) + 0.38 \times \log_2(0.38)]$

= $-[0.62 \times -0.69 + 0.38 \times -1.40]$

= $-[0.43 + 0.53]$

= 0.23

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)
+ p(ผล = No) × entropy(ผล = No)]

= 0.86 – [0.65 × 0.78 + 0.35 × 0.23]

= 0.86 – [0.51 + 0.08]

= 0.86 – 0.59

= 0.27

2) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) และแอตทริบิวต์ work interfere (Sometimes) ไปแอตทริบิวต์ coworkers(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.72 \times \log_2(0.72) + 0.28 \times \log_2(0.28)]$

= $-[0.72 \times -0.47 + 0.28 \times -1.84]$

= $-[0.34 + 0.52]$

= 0.86

entropy(ผล = Yes) = $-p(Yes) \times \log_2 p(Yes) + p(No) \times \log_2 p(No)$

= $-[0.74 \times \log_2(0.74) + 0.26 \times \log_2(0.57)]$

= $-[0.74 \times -0.43 + 0.26 \times -1.94]$

= $-[0.32 + 0.50]$

= 0.82

entropy(ผล = No) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.61 \times \log_2(0.61) + 0.39 \times \log_2(0.39)]$

$= -[0.61 \times -0.71 + 0.39 \times -1.36]$

$= -[0.43 + 0.53]$

$= 0.96$

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes) + p(ผล = No) × entropy(ผล = No)]

$= 0.86 - [0.81 \times 0.82 + 0.19 \times 0.96]$

$= 0.86 - [0.66 + 0.18]$

$= 0.86 - 0.84$

$= 0.02$

จากการคำนวณค่า IG ของแอตทริบิวต์ family history(Yes) ต่อแอตทริบิวต์ work interfere(Sometimes) ไปแอตทริบิวต์ supervisor และแอตทริบิวต์ coworkers พบว่าค่า IG ของแอตทริบิวต์ family history ต่อแอตทริบิวต์ work interfere ไปแอตทริบิวต์ supervisor มีค่ามากที่สุด (0.27) ดังนั้นจึงเลือกแอตทริบิวต์ supervisor ขึ้นมาเป็นโหนดในระดับที่ 3 ต่อจากโหนด Root ต่อจากโหนดระดับที่ 2 และทำการแตกกิ่งจากโหนดในระดับที่ 3 ออกไปจนข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกัน

การคำนวณค่าแต่ละแอตทริบิวต์ในระดับที่ 4 กับจำนวนข้อมูลทั้งหมดโดยใช้ผลลัพธ์เป็นควรไปรักษา และไม่ควรไปรักษา ดังนี้

1) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) แอตทริบิวต์ work interfere(Sometimes) และ แอตทริบิวต์ supervisor(No) ไปแอตทริบิวต์ cowoker(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.62 \times \log_2(0.62) + 0.38 \times \log_2(0.38)]$

$= -[0.62 \times -0.69 + 0.38 \times -1.40]$

$= -[0.43 + 0.53]$

$= 0.96$

entropy(ผล = Yes) $= -p(\text{Yes}) \times \log_2 p(\text{Yes}) + p(\text{No}) \times \log_2 p(\text{No})$

$= -[0.64 \times \log_2(0.64) + 0.36 \times \log_2(0.36)]$

$= -[0.64 \times -0.64 + 0.36 \times -1.47]$

= –[0.41 + 0.53]

= 0.94

entropy(ผล = No) = –p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= –[0.20 × $\log_2$(0.20) + 0.80 × $\log_2$(0.80)]

= –[0.20 × –2.32 + 0.80 × –0.32]

= –[0.46 + 0.26]

= 0.72

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)
+ p(ผล = No) × entropy(ผล = No)]

= 0.96 – [0.95 × 0.94 + 0.05 × 0.72]

= 0.96 – [0.89 + 0.03]

= 0.96 – 0.92

= 0.4

2) คำนวณค่า IG ของแอตทริบิวต์ family history(Yes) แอตทริบิวต์ work interfere(Sometimes) และแอตทริบิวต์ supervisor(Yes) ไปแอตทริบิวต์ cowoker(Yes และ No) จากข้อมูลสามารถคำนวณค่า IG ได้ดังนี้

entropy (parent) = –p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= –[0.77 × $\log_2$(0.77) + 0.23 × $\log_2$(0.23)]

= –[0.77 × –0.38 + 0.23 × –2.12]

= –[0.29 + 0.49]

= 0.78

entropy(ผล = Yes) = –p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= –[0.81 × $\log_2$(0.81) + 0.19 × $\log_2$(0.19)]

= –[0.81 × –0.30 + 0.19 × –2.40]

= –[0.24 + 0.46]

= 0.70

entropy(ผล = No) = –p(Yes) × $\log_2$ p(Yes) + p(No) × $\log_2$ p(No)

= –[0.65 × $\log_2$(0.65) + 0.35 × $\log_2$(0.35)]

= –[0.65 × –0.62 + 0.35 × –1.51]

= –[0.40 + 0.53]

= 0.93

IG(parent, child) = entropy (parent) – [p(ผล = Yes) × entropy(ผล = Yes)
+ p(ผล = No) × entropy(ผล = No)]

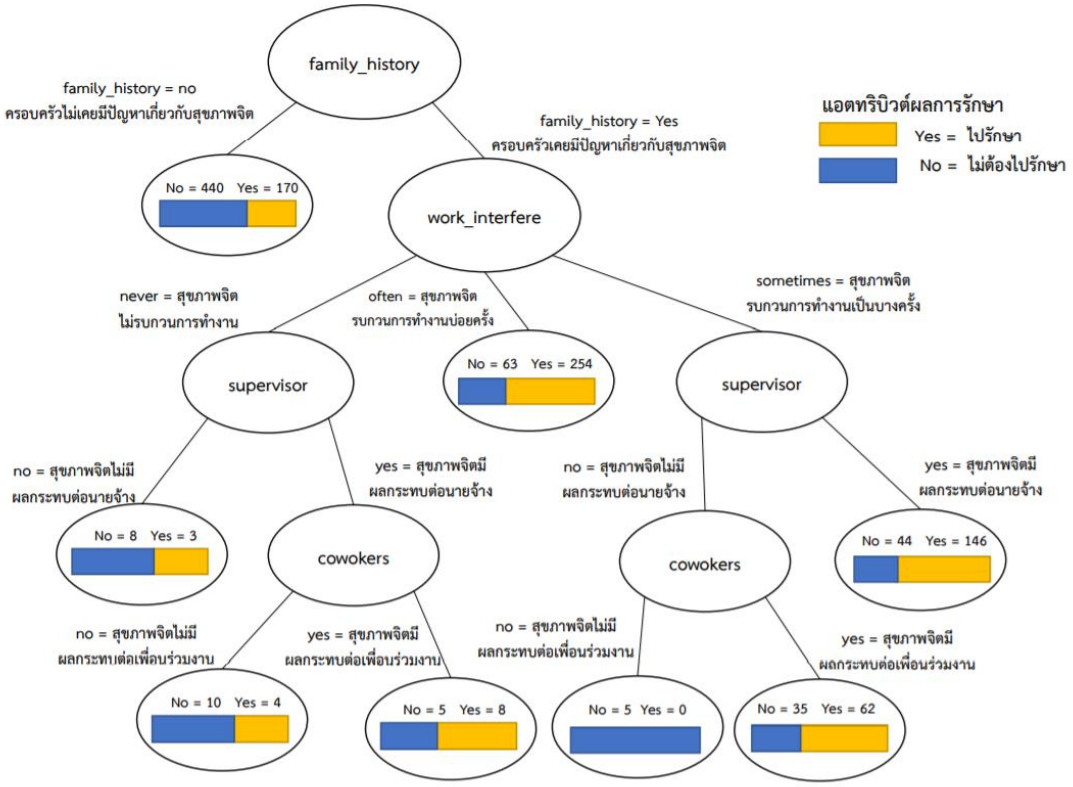
= 0.78 – [0.73 × 0.70 + 0.27 × 0.93]

= 0.78 – [0.51 + 0.25]

= 0.78 – 0.76

= –0.02

จากการคำนวณค่า IG ของแอตทริบิวต์ family history(Yes) แอตทริบิวต์ work interfere (Sometimes) แอตทริบิวต์ supervisor(Yes) แอตทริบิวต์ cowoker(Yes และ No) มีความสัมพันธ์กับผลการรักษามากที่สุดอยู่ที่ 0.4 และ 0.2 ตามลำดับ และพบว่าข้อมูลในแต่ละโหนดมีคลาสคำตอบเดียวกันแล้ว คือ ผลลัพธ์เป็นควรไปรักษาและไม่ควรไปรักษาตามภาพที่ 3.9

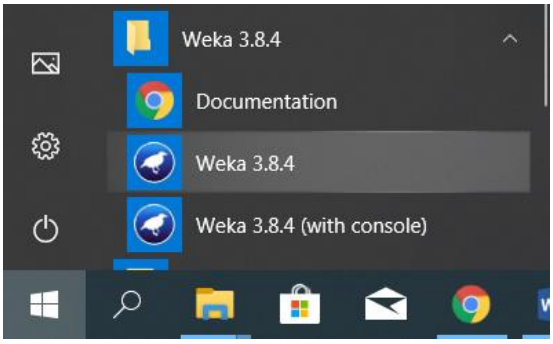


ภาพที่ 3.9 แสดงผลลัพธ์แผนภาพโมเดลต้นไม้ตัดสินใจ จากการคำนวณด้วยมือ

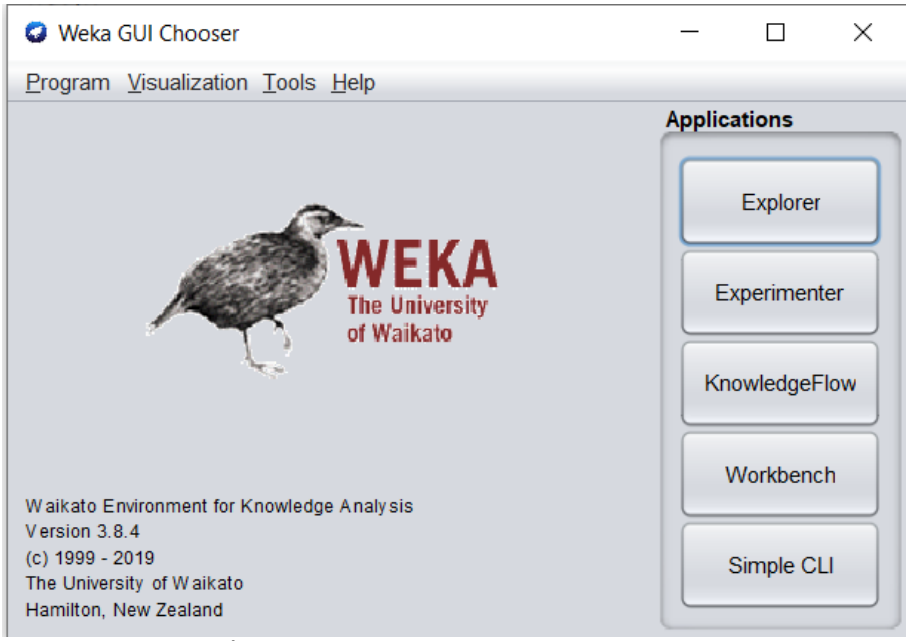
3.1.5 การประเมินผล (Evaluation) เป็นขั้นตอนก่อนนำผลลัพธ์ที่ได้จากขั้นตอนที่ 4 ไปใช้งาน ด้วยการวัดประสิทธิผลของผลลัพธ์ที่ได้กับวัตถุประสงค์ที่ตั้งไว้ในขั้นตอนแรก ว่ามีนัยสำคัญหรือความน่าเชื่อถือมากน้อยเพียงใด ด้วยการประเมินผลจากโปรแกรมว่าถูกต้องหรือไม่

ผู้วิเคราะห์ข้อมูลได้ทำการทดสอบโมเดล เพื่อวัดประสิทธิภาพที่ตรงกับความต้องการ ซึ่งการวัดประสิทธิภาพด้วยวิธี Self-Consistency Test เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพ เพื่อดูแนวโน้มของโมเดลที่สร้างขึ้น และเมื่อนำข้อมูลมาทดสอบ (Testing data) กับโปรแกรมที่ผู้วิเคราะห์เลือกมาทดสอบกับข้อมูลที่ผ่านการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining จากการสร้างโมเดล Decision Tree จึงนำข้อมูลดังกล่าว มาทดสอบกับโปรแกรม Weka เวอร์ชัน 3.8.4 ซึ่งมีขั้นตอนการทำงาน ดังนี้

ขั้นตอนที่ 1 เปิดโปรแกรม Weka 3.8.4

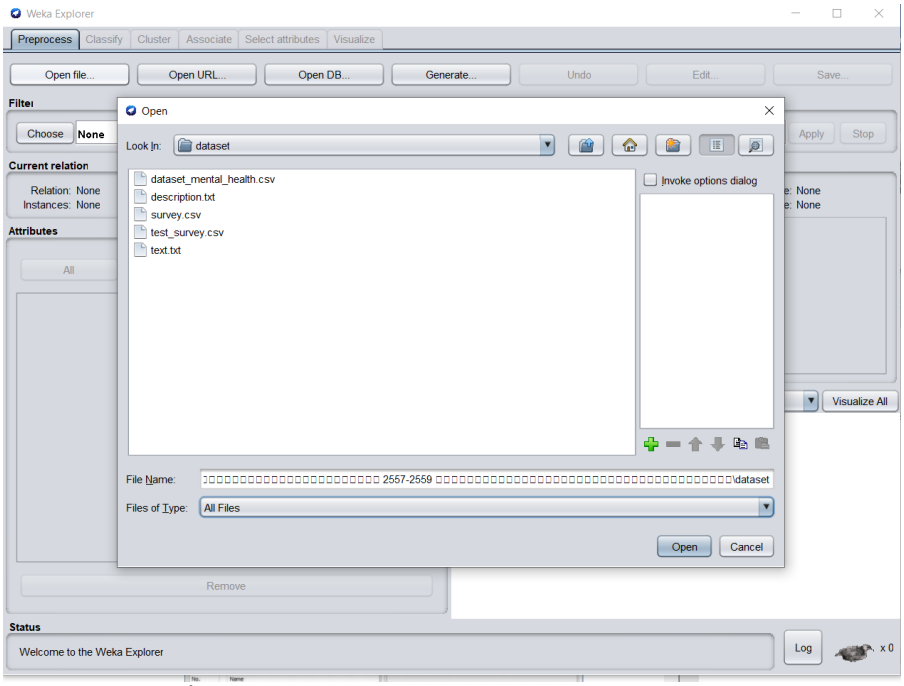


ภาพที่ 3.10 แสดงการเปิดโปรแกรม Weka 3.8.4

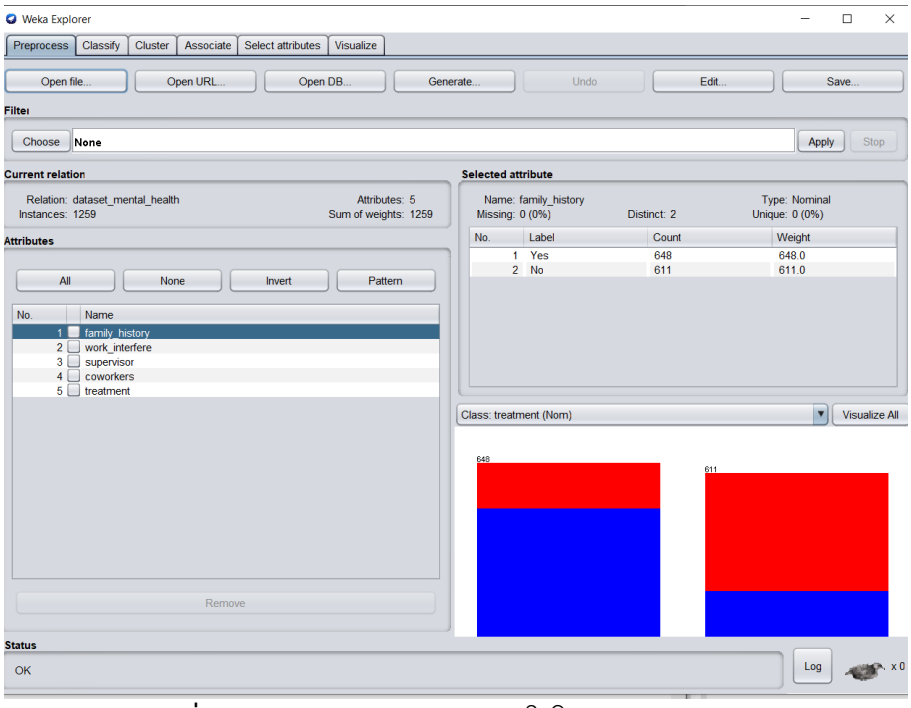


ภาพที่ 3.11 แสดงการเข้าหน้าจอโปรแกรม Weka 3.8.4

ขั้นตอนที่ 2 นำเข้าข้อมูลที่ได้จัดเตรียมไว้ โดยเลือกที่ Application >>Explorer>>Open file เลือกไฟล์ข้อมูลที่ต้องการนำมาทดสอบตามภาพที่ 3.12 และหลังจากนั้นโปรแกรมแสดงหน้าจอข้อมูล ตามภาพที่ 3.13

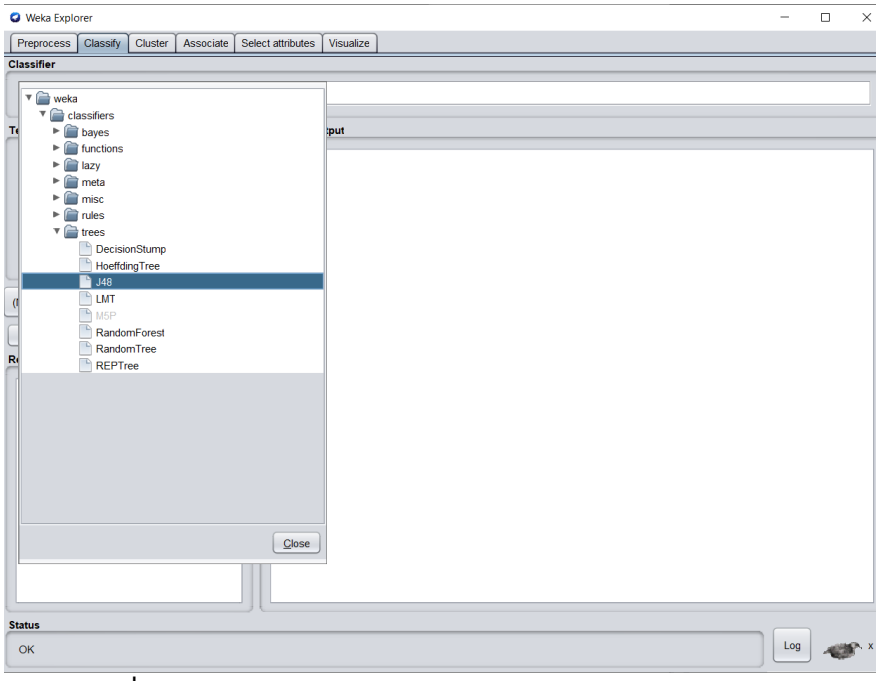


ภาพที่ 3.12 แสดงการนำเข้าข้อมูลเข้าในโปรแกรม Weka 3.8.4

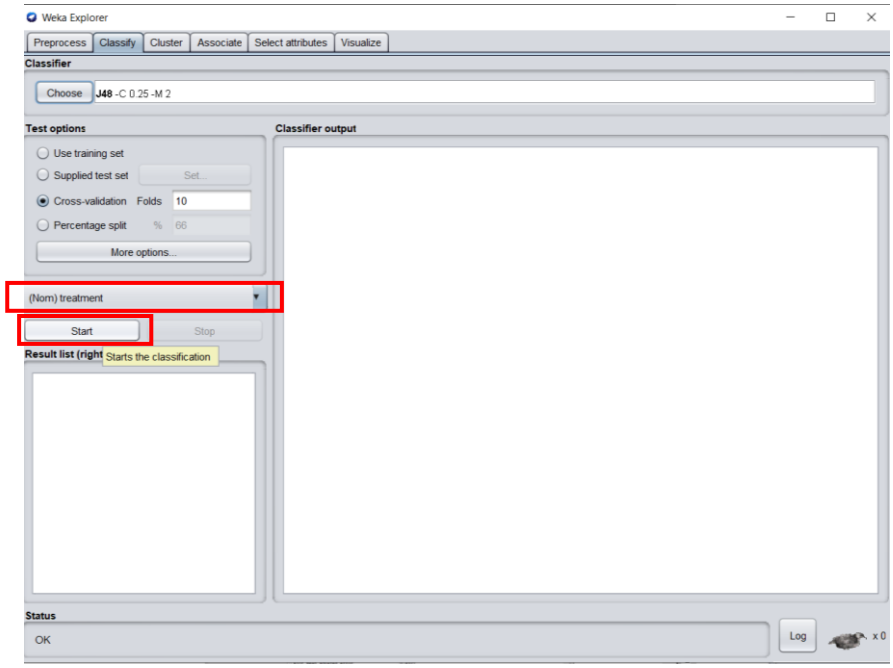


ภาพที่ 3.13 แสดงการนำเข้าข้อมูลเข้าในโปรแกรม Weka 3.8.4

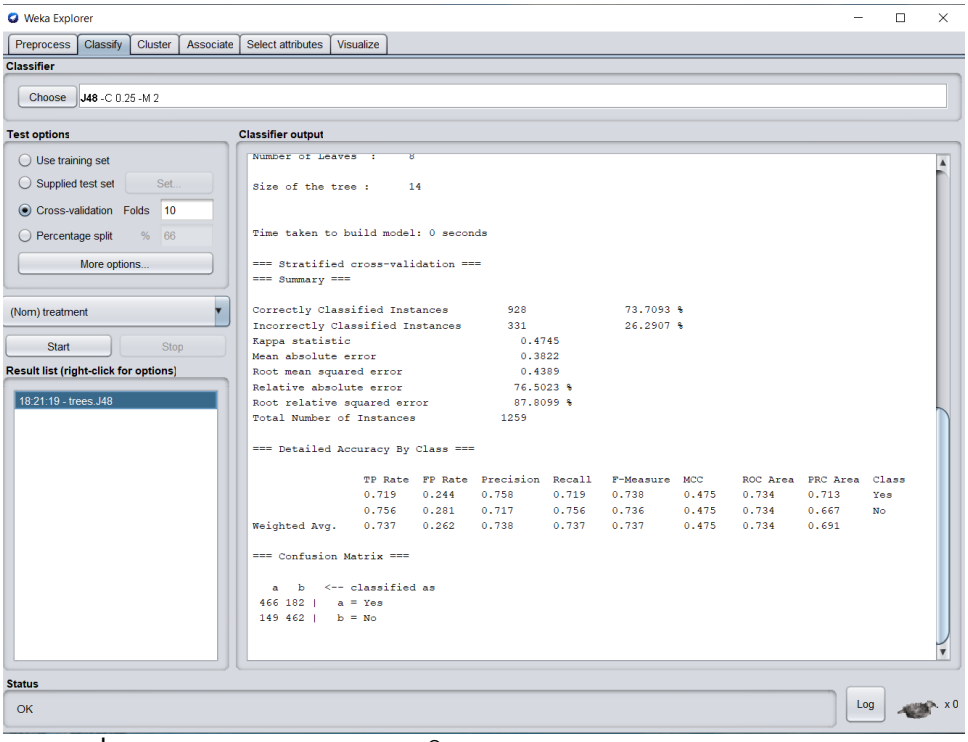
ขั้นตอนที่ 3 ดำเนินการเลือกเทคนิคการจัดกลุ่มข้อมูลแบบ Decision Tree โดยเลือกที่ Classification>>Choose>>tree และเลือกรูปแบบเป็น J8 ตามภาพที่ 3.14 จากนั้นเลือกตัวบ่งชี้ในที่นี้ใช้เป็นผลการรักษา (treatment) จากนั้นกดปุ่ม Start ตามภาพที่ 3.15 จะแสดงผลลัพธ์ที่ได้ตามภาพที่ 3.16



ภาพที่ 3.14 แสดงการเลือกเทคนิคการจัดกลุ่มข้อมูลแบบ Decision Tree

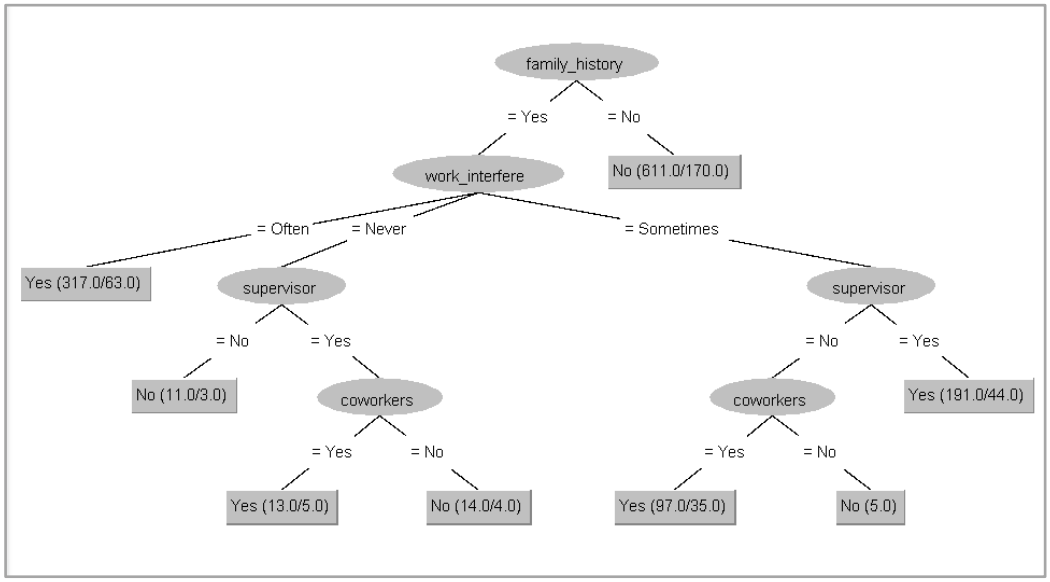


ภาพที่ 3.15 แสดงการเลือกเทคนิคการจัดกลุ่มข้อมูลแบบ Decision Tree: J48



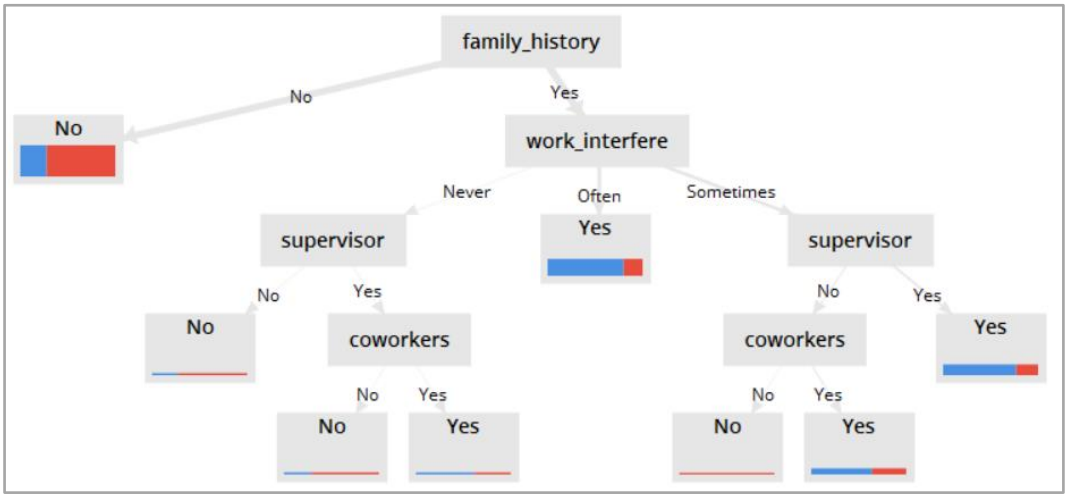
ภาพที่ 3.16 แสดงหน้าจอผลลัพธ์ของโมเดลการจัดกลุ่มข้อมูลแบบ Decision Tree: J48

จากผลลัพธ์การทดลองพบว่าเทคนิค Decision Tree: J48 ให้ผลลัพธ์การจำแนกประเภทผลการรักษา ไปรักษา(Yes) และไม่ไปรักษา(No) มีความถูกต้องถึง 73.71% แสดงผลลัพธ์แผนภาพโมเดลต้นไม้ตัดสินใจที่มีกิ่งแตกออกมา ดังภาพที่ 3.17

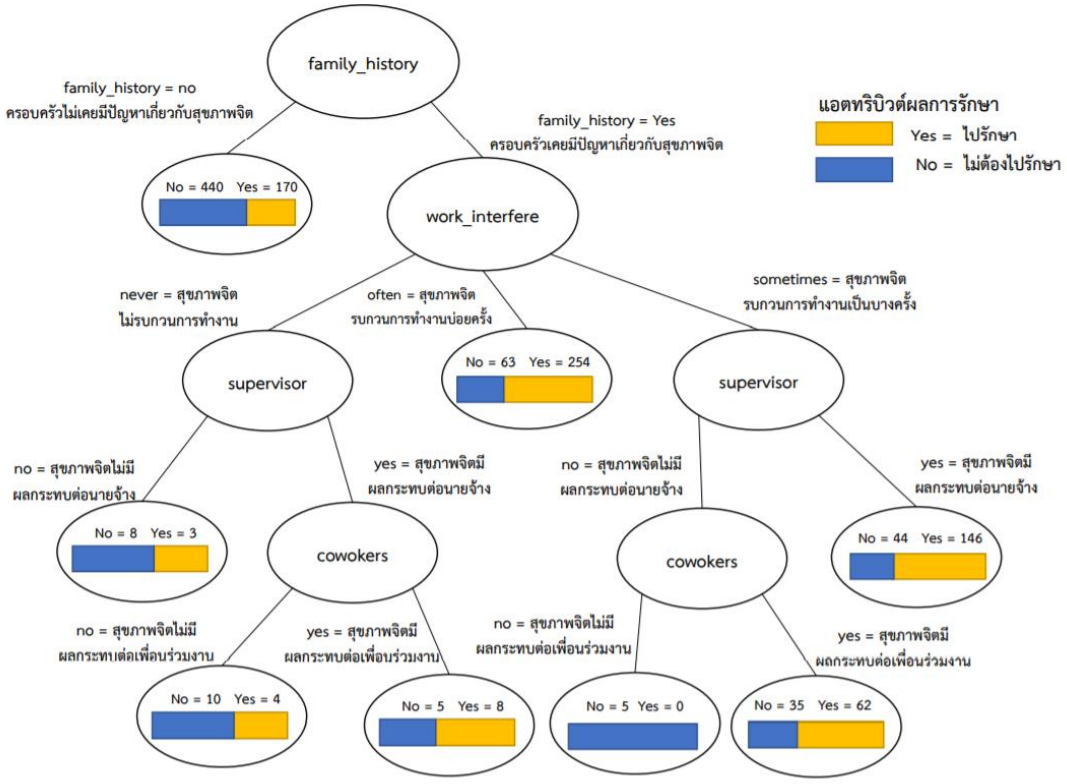


ภาพที่ 3.17 แสดงผลลัพธ์แผนภาพโมเดลต้นไม้ตัดสินใจ ในโปรแกรม Weka 3.8.4

ทางผู้วิเคราะห์ข้อมูลยังได้นำโมเดลของข้อมูลที่เลือกใช้ มาเปรียบเทียบกับโมเดลที่สร้างด้วยโปรแกรม RapidMiner Studio 9.5.1 เพื่อตรวจสอบความแม่นยำของโมเดล ซึ่งผู้วิเคราะห์ข้อมูลพบว่าได้ผลลัพธ์รูปแบบแผนภาพโมเดลที่ตรงกันสูง



ภาพที่ 3.18 แสดงผลลัพธ์แผนภาพโมเดลต้นไม้ตัดสินใจ ของโปรแกรม RapidMiner Studio



ภาพที่ 3.19 แสดงผลลัพธ์แผนภาพโมเดลต้นไม้ตัดสินใจ จากการคำนวณด้วยมือ

จากผลลัพธ์การคำนวณค่า IG ของโมเดลปัญหาสุขภาพจิตในการทำงานในแต่ละโหนด และจากการทำการทดสอบโมเดล เพื่อวัดประสิทธิภาพที่ตรงกับความต้องการด้วยวิธี Self-Consistency Test เพื่อดูแนวโน้มของโมเดลที่สร้างขึ้นจากโปรแกรม Weka เวอร์ชัน 3.8.4 และเมื่อนำไปเปรียบเทียบกับโปรแกรม RapidMiner Studio เวอร์ชัน 9.5.1 ทางผู้วิเคราะห์ข้อมูลพบว่าทั้ง 3 โมเดลได้ผลลัพธ์ความแม่นยำของโมเดลที่ตรงกันสูง ดังนั้นทางผู้วิเคราะห์ข้อมูลจึงได้เลือกโมเดลจากวิธีการคำนวณมือที่มีความน่าเชื่อถือนี้ไปใช้งานต่อไป

จากผลลัพธ์การสร้างโมเดลด้วยเทคนิค Decision Tree: J48 ในโปรแกรม Weka 3.8.4 ได้สร้างกฎจากการจำแนกกลุ่มต้นไม้การตัดสินใจแบบ Decision Tree มีกิ่งแตกออกมา ดังภาพที่ 3.19

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

family_history = Yes
|   work_interfere = Often: Yes (317.0/63.0)
|   work_interfere = Never
|   |   supervisor = No: No (11.0/3.0)
|   |   supervisor = Yes
|   |   |   coworkers = Yes: Yes (13.0/5.0)
|   |   |   coworkers = No: No (14.0/4.0)
|   work_interfere = Sometimes
|   |   supervisor = No
|   |   |   coworkers = Yes: Yes (97.0/35.0)
|   |   |   coworkers = No: No (5.0)
|   |   supervisor = Yes: Yes (191.0/44.0)
family_history = No: No (611.0/170.0)

Number of Leaves  :     8

Size of the tree :      14
```

ภาพที่ 3.20 แสดงผลลัพธ์กฎต้นไม้การตัดสินใจ ของโปรแกรม Weka 3.8.4

ซึ่งผู้วิเคราะห์ข้อมูลพบว่าเมื่อนำมาเปรียบเทียบกับกฎจากการจำแนกกลุ่มแบบ Decision Tree ในโปรแกรม RapidMiner Studio ได้ผลลัพธ์รูปแบบกฎจากการจำแนกกลุ่มที่ตรงกันสูง

Tree

```
family_history = No: No {Yes=170, No=441}
family_history = Yes
|   work_interfere = Never
|   |   supervisor = No: No {Yes=3, No=8}
|   |   supervisor = Yes
|   |   |   coworkers = No: No {Yes=4, No=10}
|   |   |   coworkers = Yes: Yes {Yes=8, No=5}
|   work_interfere = Often: Yes {Yes=254, No=63}
|   work_interfere = Sometimes
|   |   supervisor = No
|   |   |   coworkers = No: No {Yes=0, No=5}
|   |   |   coworkers = Yes: Yes {Yes=62, No=35}
|   |   supervisor = Yes: Yes {Yes=147, No=44}
```

ภาพที่ 3.21 แสดงผลลัพธ์กฎต้นไม้การตัดสินใจ ในโปรแกรม RapidMiner Studio

ดังนั้นผู้วิเคราะห์ข้อมูลจะใช้เทคนิคของการจำแนกกลุ่มแบบ Decision Tree: J48 มาใช้ในการศึกษา เนื่องจากให้ผลลัพธ์ของกฎที่สามารถทำนายได้จำนวน 8 กฎ ซึ่งสามารถนำไปใช้ในการแบ่งกลุ่มได้ตามเงื่อนไขได้ชัดเจน และสามารถนำกฎที่ได้ไปวิเคราะห์กฎต่อไปได้ โดยสามารถจำแนกกฎได้ ดังนี้

กฎข้อที่ 1 IF Family_history=No Then Treatment=No หมายความว่า ถ้าครอบครัวไม่มีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ไม่ต้องไปรักษา

กฎข้อที่ 2 IF Family_history=Yes And Work_interfere=Often Then Treatment=Yes หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตรบกวนการทำงานอยู่บ่อยครั้ง ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ต้องไปรักษา

กฎข้อที่ 3 IF Family_history=Yes And Work_interfere=Never And Supervisor=No Then Treatment=No หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตไม่รบกวนการทำงาน และไม่มีผลกระทบต่อนายจ้าง ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ไม่ต้องไปรักษา

กฎข้อที่ 4 IF Family_history=Yes And Work_interfere=Never And Supervisor=Yes And Coworkers=Yes Then Treatment=Yes หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตไม่รบกวนการทำงาน แต่สุขภาพจิตมีผลกระทบต่อนายจ้างและเพื่อนร่วมงาน ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ต้องไปรักษา

กฎข้อที่ 5 IF Family_history=Yes And Work_interfere=Never And Supervisor=Yes And Coworkers=No Then Treatment=No หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตไม่รบกวนการทำงาน สุขภาพจิตมีผลกระทบต่อนายจ้าง แต่ไม่มีผลกระทบต่อเพื่อนร่วมงาน ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ไม่ต้องไปรักษา

กฎข้อที่ 6 IF Family_history=Yes And Work_interfere=Sometimes And Supervisor=No And Coworkers=Yes Then Treatment=Yes หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตรบกวนการทำงานเป็นบางครั้ง สุขภาพจิตไม่มีผลกระทบต่อนายจ้าง แต่มีผลกระทบต่อเพื่อนร่วมงาน ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ต้องไปรักษา

กฎข้อที่ 7 IF Family_history=Yes And Work_interfere=Sometimes And Supervisor=No And Coworkers=No Then Treatment=No หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตรบกวนการทำงานเป็นบางครั้ง แต่สุขภาพจิตไม่มีผลกระทบต่อนายจ้างและเพื่อนร่วมงาน ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ไม่ต้องไปรักษา

กฎข้อที่ 8 IF Family_history=Yes And Work_interfere=Sometimes And Supervisor=Yes Then Treatment=Yes หมายความว่า ถ้าครอบครัวมีประวัติเกี่ยวกับปัญหาด้านสุขภาพจิต และสุขภาพจิตรบกวนการทำงานเป็นบางครั้ง และสุขภาพจิตมีผลกระทบต่อนายจ้าง ผลการพิจารณาพบว่า ผลการรักษาจะมีผลลัพธ์เป็น ต้องไปรักษา

หลังจากผู้วิเคราะห์ข้อมูลเลือกการทดสอบประสิทธิภาพของ Model ด้วยวิธี Self-Consistency Test หรือเรียกว่า Use Training Set เป็นวิธีการที่นำข้อมูลที่ใช้ในการสร้างโมเดล (model) และข้อมูลที่ใช้ในการทดสอบโมเดลเป็นข้อมูลชุดเดียวกัน คือข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 ที่ได้ทำการคัดเลือกมาทั้งหมด ซึ่งผู้วิเคราะห์ข้อมูลเลือกโปรแกรมที่ใช้นำเสนอ คือ โปรแกรม Weka 3.8.4 พบว่าการทดสอบประสิทธิภาพโมเดล Decision Tree (J48) พิจารณาได้ว่า โมเดลที่ถูกสร้างขึ้น มีค่าความถูกต้องเฉลี่ยในทุกโมเดลเท่ากับ 73.71% มีค่าการทำนายข้อมูลไม่ถูกต้องเท่ากับ 26.29% และมีค่าความคลาดเคลื่อนเท่ากับ 0.4389 และเมื่อพิจารณาส่วนค่า Confusion Matrix ในภาพที่ 3.21 พบว่าการหาค่าของข้อมูลค่าจริง กับจำนวนข้อมูลจากการทำนาย แบ่งตามประเภทของผลการรักษา ต้องไปรักษา (Yes) และไม่ต้องไปรักษา (No) และนำมาหาค่าเฉลี่ยรวมของทุก class ได้ค่าเฉลี่ยรวมเท่ากับ 0.74 มีผลลัพธ์ตรงกันอยู่ในระดับค่อนข้างดี สามารถนำโมเดลไปใช้งานได้

ดังนั้น ผู้วิเคราะห์ข้อมูลจึงเลือกใช้โมเดล Decision Tree: J48 ในการนำไปหาแนวทางการวางแผนการจัดการและรับมือต่อปัจจัยเสี่ยงต่าง ๆ ของผู้ที่ประสบปัญหาสุขภาพจิตในการทำงานในปัจจุบัน และยังเป็นปัจจัยชี้นำประสิทธิภาพการทำงานขององค์กรและภาครัฐต่าง ๆ ได้อีกด้วย เพราะมีค่าความถูกต้องของโมเดล และค่าเฉลี่ย Confusion Matrix จากผลการรักษาต้องไปรักษา (Yes) และไม่ต้องไปรักษา (No) อยู่ในระดับที่ค่อนข้างดี

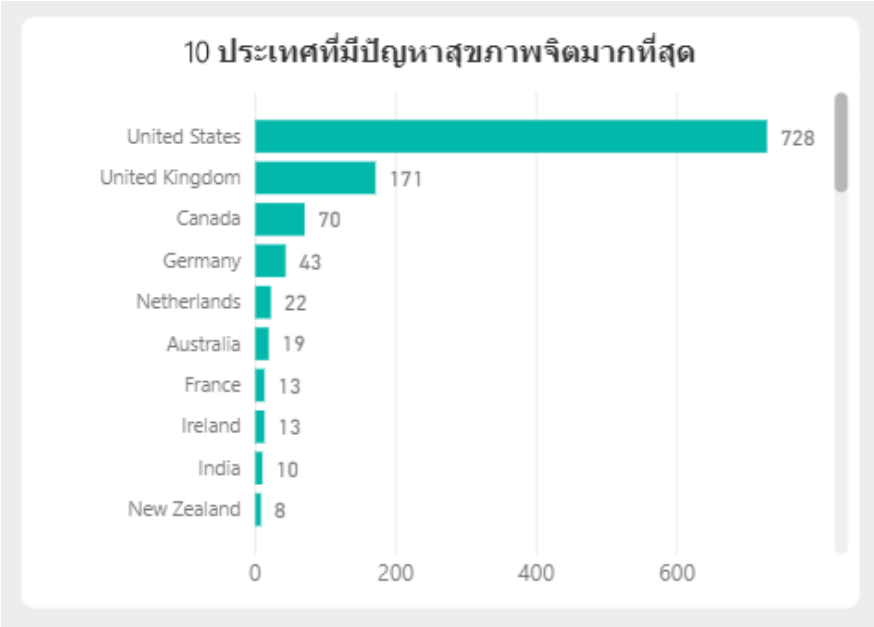
```
=== Summary ===

Correctly Classified Instances         928               73.7093 %
Incorrectly Classified Instances       331               26.2907 %
Kappa statistic                          0.4745
Mean absolute error                      0.3822
Root mean squared error                  0.4389
Relative absolute error                 76.5023 %
Root relative squared error             87.8099 %
Total Number of Instances             1259

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.719    0.244    0.758      0.719    0.738      0.475    0.734     0.713     Yes
                 0.756    0.281    0.717      0.756    0.736      0.475    0.734     0.667     No
Weighted Avg.    0.737    0.262    0.738      0.737    0.737      0.475    0.734     0.691

=== Confusion Matrix ===

   a   b   <-- classified as
 466 182 |   a = Yes
 149 462 |   b = No
```

ภาพที่ 3.22 แสดงผลลัพธ์จากการจำแนกกลุ่มแบบ Decision Tree: J48 ในโปรแกรม Weka 3.8.4

3.1.6 เผยแพร่ผลการวิเคราะห์ (Deployment) ขั้นตอนการนำเอาผลลัพธ์ที่ได้ไปใช้งานเป็นการทั่วไป อาจจัดทำเป็นรูปแบบของรายงาน (Report) หรือแผนภาพ (Dashboard) ที่พร้อมให้ฝ่ายต่าง ๆ นำไปใช้ประโยชน์ในการวางแผน กำหนดกลยุทธ์ และดำเนินการต่าง ๆ ในทางธุรกิจ

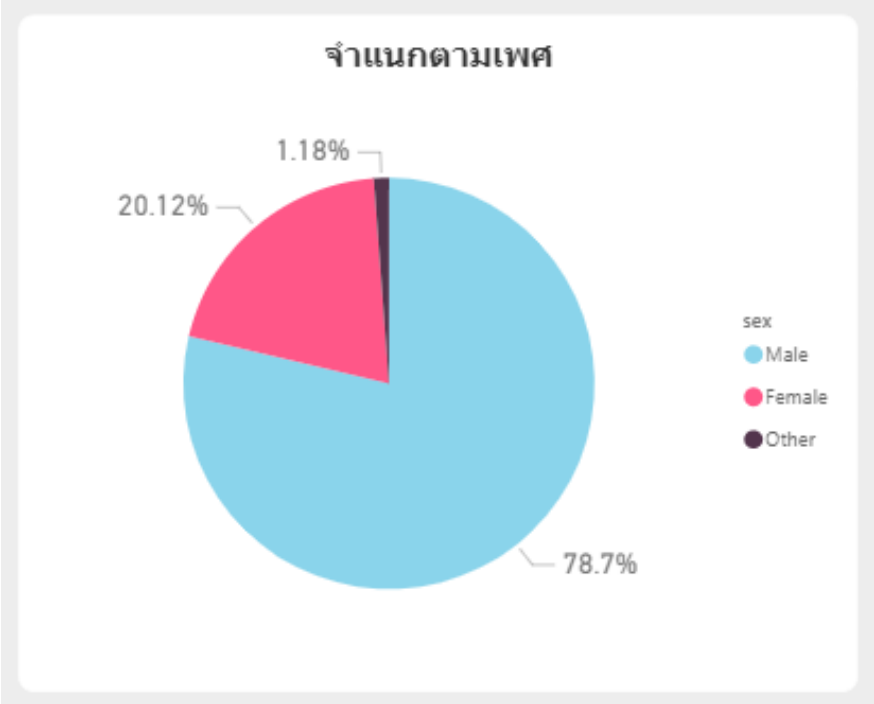
ผู้วิเคราะห์ข้อมูลนำผลข้อมูลที่ทำการวิเคราะห์แล้ว แสดงผลข้อมูลบน Web Browser โดยใช้ชุดคำสั่ง HTML และ CSS3 ร่วมกับการนำเสนอข้อมูลแบบ Visualization ด้วยการแสดงผลข้อมูลในรูปแบบของภาพโดยใช้โปรแกรม Power BI ซึ่งทางผู้วิเคราะห์ข้อมูลได้ยกตัวอย่างการจัดทำเป็นรูปแบบของรายงาน (Report) หรือแผนภาพ (Dashboard) กับข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557 – 2559 ดังนี้

1) แสดงผลการสรุปข้อมูลปัญหาสุขภาพจิตในการทำงานในภาพรวมตั้งแต่ปี 2557 – 2559 แสดงประเทศที่มีปัญหาสุขภาพจิตสูงสุด 10 อันดับ



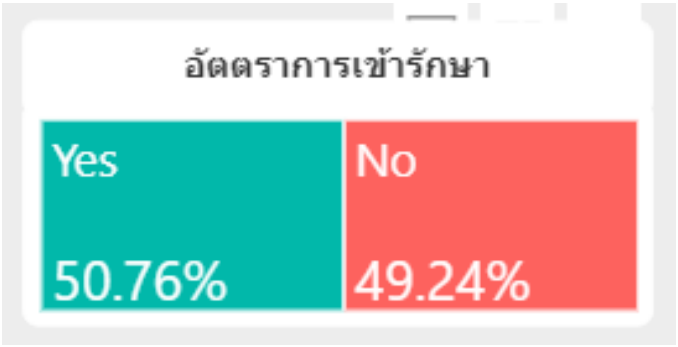
ภาพที่ 3.23 แสดงข้อมูลการสรุปข้อมูลปัญหาสุขภาพจิตในการทำงานในภาพรวม

2) แสดงผลการสรุปข้อมูลปัญหาสุขภาพจิตในการทำงานในภาพรวมจำแนกตามเพศสภาพจริง ดังนี้ เพศชาย เพศหญิง และเพศอื่น ๆ ที่ไม่สามารถระบุเพศได้



ภาพที่ 3.24 แสดงข้อมูลการสรุปข้อมูลปัญหาสุขภาพจิตในการทำงานในภาพรวมจำแนกตามเพศสภาพจริง

3) แสดงผลการสรุปข้อมูลอัตราการเข้ารับการรักษาปัญหาสุขภาพจิตในการทำงานในภาพรวมทั้งหมด

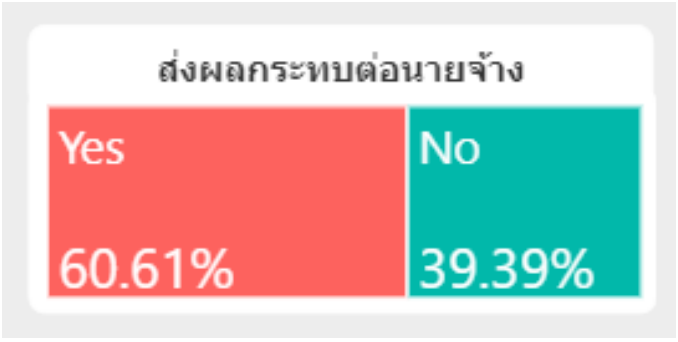


ภาพที่ 3.25 แสดงข้อมูลการสรุปอัตราการเข้ารับการรักษาปัญหาสุขภาพจิตในการทำงาน

4) แสดงผลการสรุปข้อมูลอัตราครอบครัวที่มีปัญหาสุขภาพจิตในภาพรวมทั้งหมด

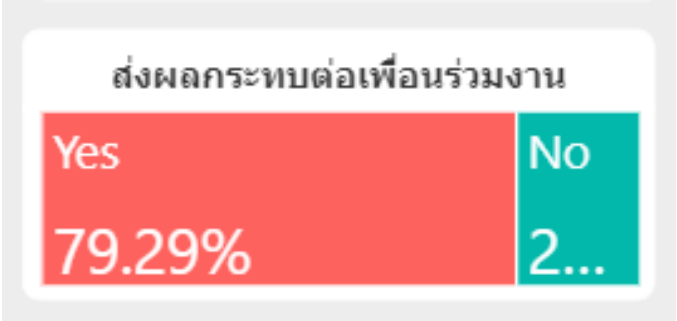
ภาพที่ 3.26 แสดงข้อมูลการสรุปอัตราครอบครัวที่มีปัญหาสุขภาพจิต

5) แสดงผลการสรุปข้อมูลอัตราปัญหาสุขภาพจิตมีผลกระทบต่อนายจ้างในภาพรวมทั้งหมด



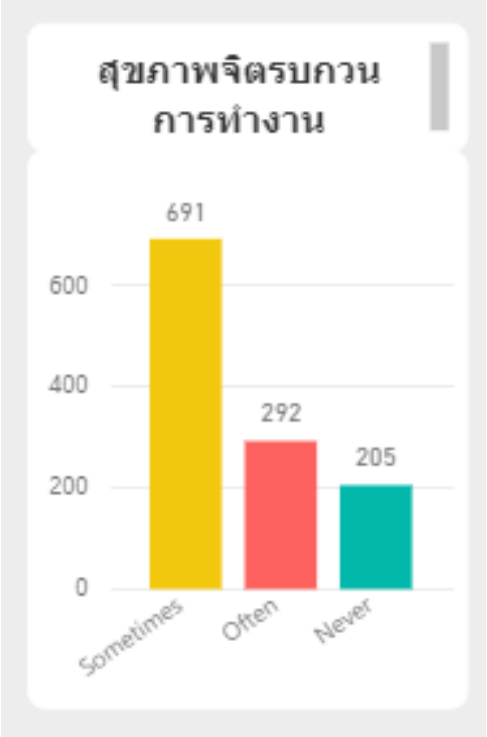
ภาพที่ 3.27 แสดงข้อมูลการสรุปอัตราปัญหาสุขภาพจิตมีผลกระทบต่อนายจ้าง

6) แสดงผลการสรุปข้อมูลอัตราปัญหาสุขภาพจิตมีผลกระทบต่อเพื่อนร่วมงานในภาพรวมทั้งหมด



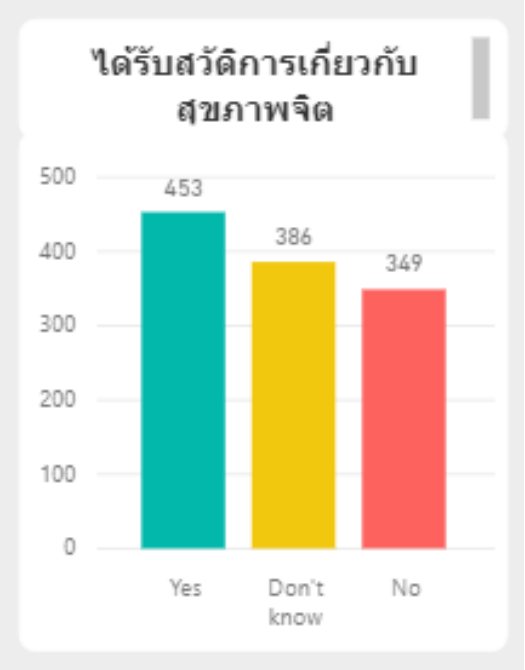
ภาพที่ 3.28 แสดงข้อมูลการสรุปอัตราปัญหาสุขภาพจิตมีผลกระทบต่อเพื่อนร่วมงาน

7) แสดงผลการสรุปข้อมูลอัตราปัญหาสุขภาพจิตรบกวนการทำงาน



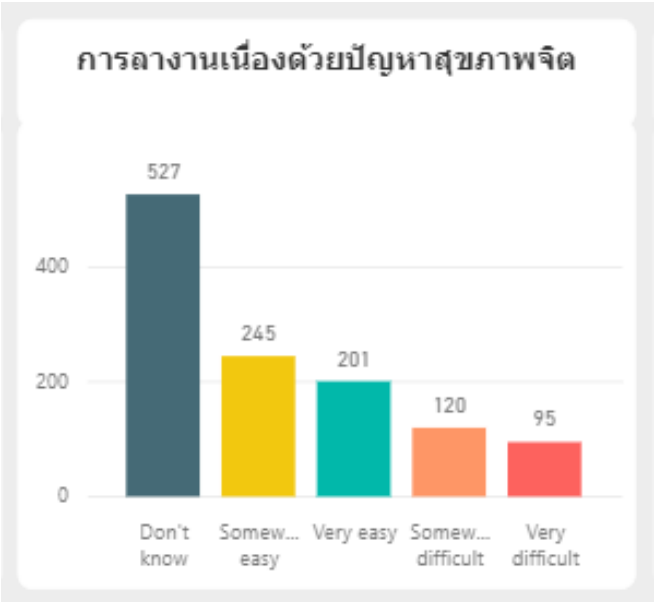
ภาพที่ 3.29 แสดงข้อมูลการสรุปอัตราปัญหาสุขภาพจิตรบกวนการทำงาน

8) แสดงผลการสรุปข้อมูลอัตราการได้รับสวัสดิการเกี่ยวกับสุขภาพจิตในการทำงาน



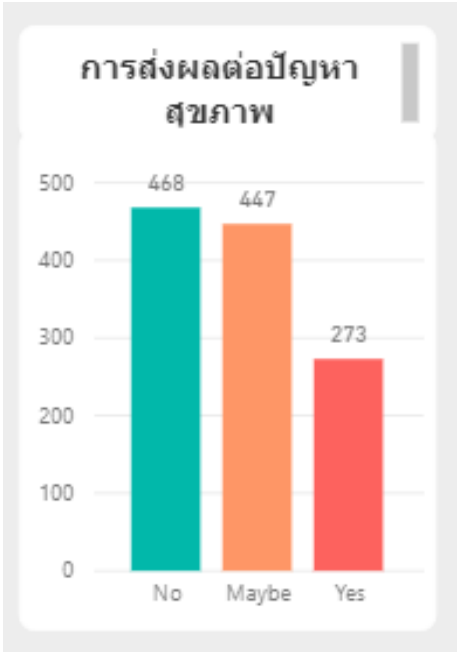
ภาพที่ 3.30 แสดงข้อมูลการสรุปอัตราการได้รับสวัสดิการเกี่ยวกับสุขภาพจิตในการทำงาน

9) แสดงผลการสรุปข้อมูลอัตราการลางานเนื่องด้วยปัญหาสุขภาพจิต



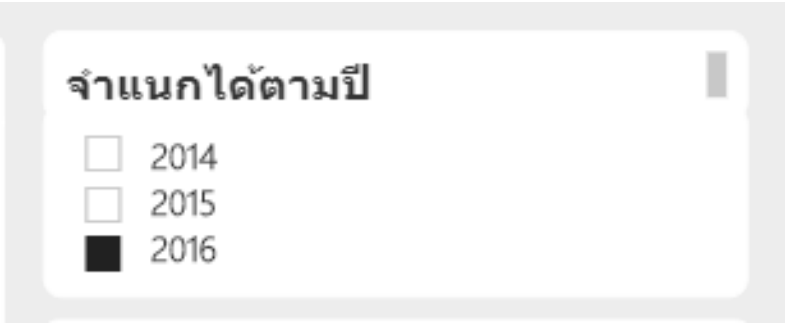
ภาพที่ 3.31 แสดงข้อมูลการสรุปอัตราการลางานเนื่องด้วยปัญหาสุขภาพจิต

10) แสดงผลการสรุปข้อมูลอัตราปัญหาสุขภาพจิตส่งผลต่อปัญหาสุขภาพ



ภาพที่ 3.32 แสดงข้อมูลการสรุปอัตราปัญหาสุขภาพจิตส่งผลต่อปัญหาสุขภาพ

11) โดยกราฟทั้งหมดนี้สามารถกรองข้อมูล ออกเป็นรายปีได้ ตั้งแต่ปี 2014 – 2016

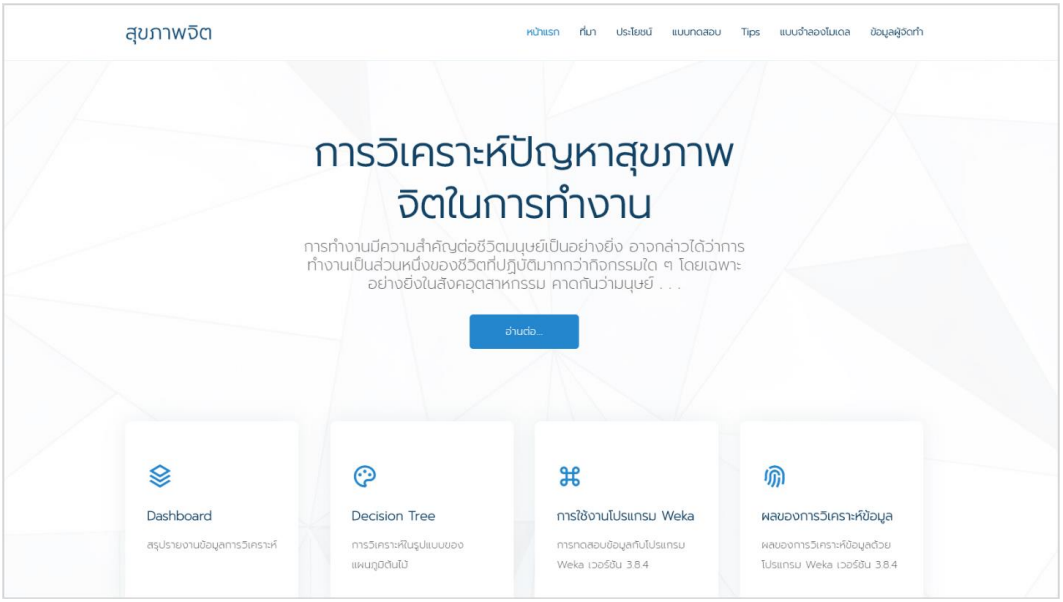


ภาพที่ 3.33 แสดงตัวกรองข้อมูลออกเป็นรายปี

3.2 การออกแบบเว็บไซต์

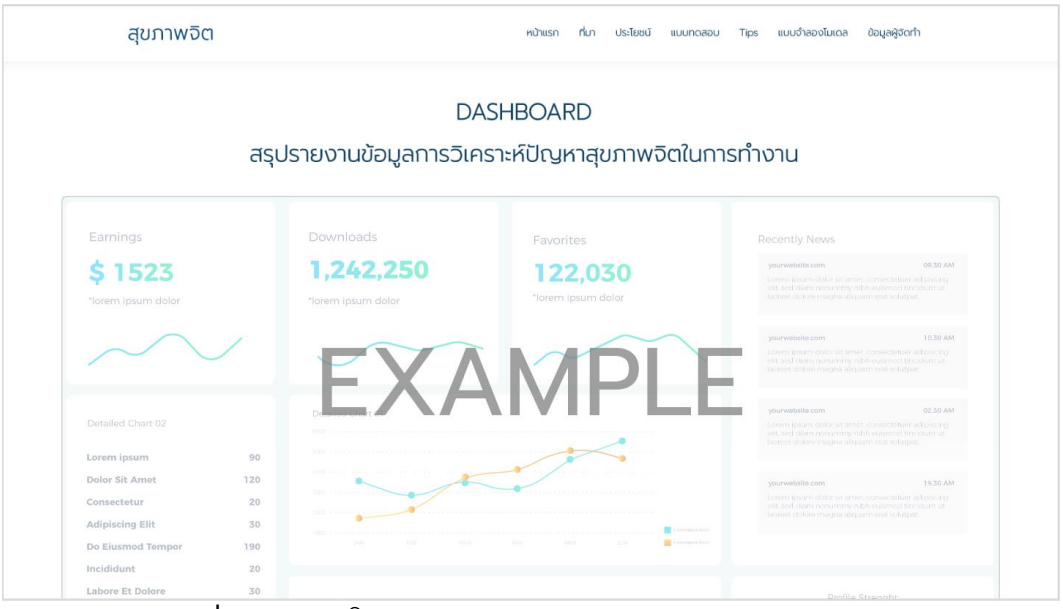
3.2.1 การออกแบบ Wireframe หน้าจอเว็บไซต์

1) หน้าแรกของเว็บไซต์ แสดงเมนูต่าง ๆ ของหน้าเว็บและเนื้อหาเกี่ยวกับปัญหาสุขภาพจิตในการทำงาน



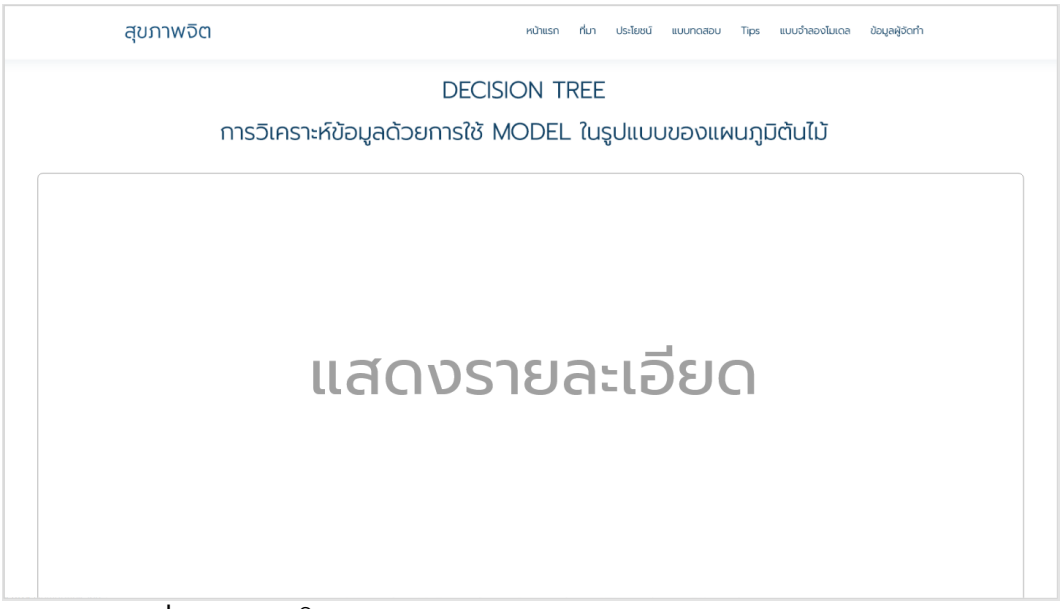
ภาพที่ 3.34 แสดงโครงร่างการออกแบบหน้าแรกของเว็บไซต์

2) หน้าแสดงผลการวิเคราะห์ข้อมูลปัญหาสุขภาพจิตในการทำงาน



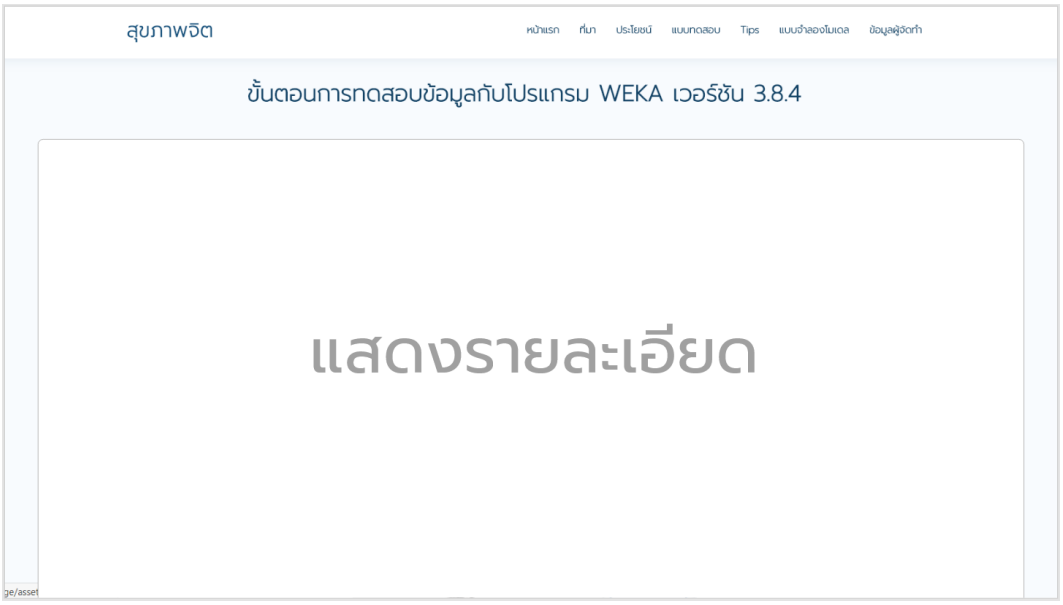
ภาพที่ 3.35 แสดงโครงร่างการออกแบบหน้าแสดงผลการวิเคราะห์ข้อมูล

3) หน้าแสดงผลการวิเคราะห์ข้อมูลด้วย Decision Tree



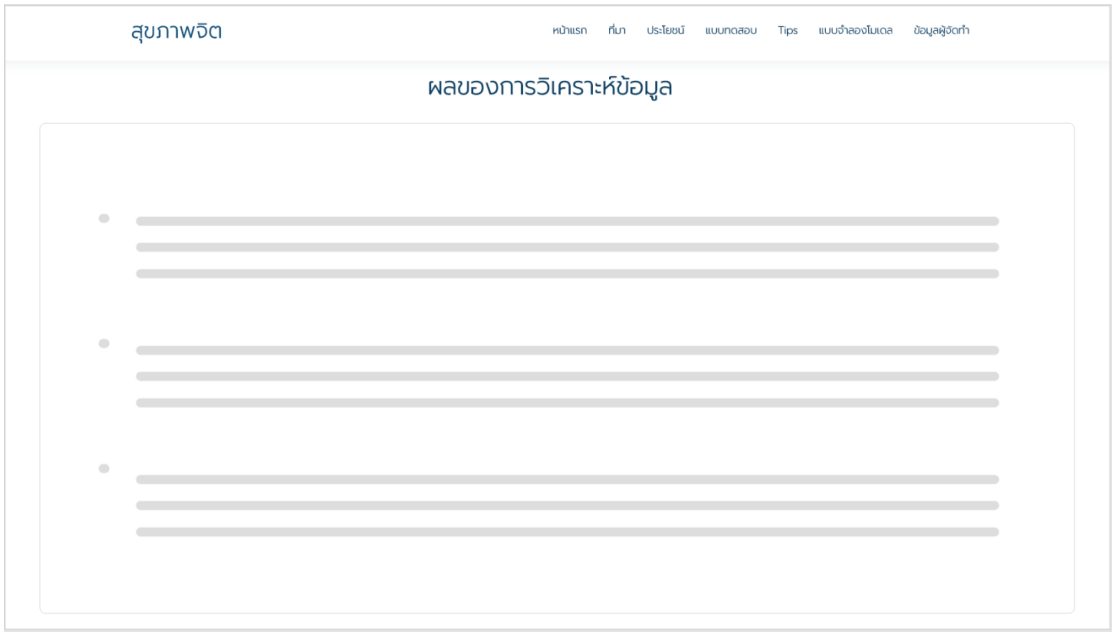
ภาพที่ 3.36 แสดงโครงร่างการออกแบบหน้าแสดงผลการวิเคราะห์ข้อมูลด้วย Decision Tree

4) หน้าแสดงผลข้อมูลขั้นตอนการทดสอบข้อมูลกับโปรแกรม Weka เวอร์ชัน 3.8.4



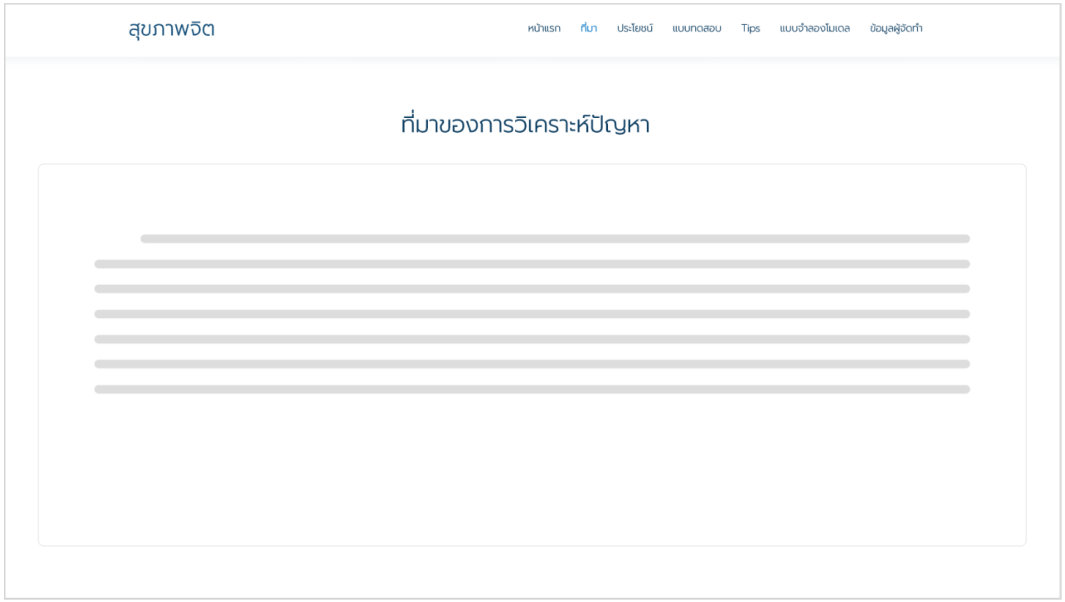
ภาพที่ 3.37 แสดงโครงร่างการออกแบบหน้าแสดงผลข้อมูลขั้นตอนการทดสอบข้อมูลกับโปรแกรม Weka เวอร์ชัน 3.8.4

5) หน้าแสดงผลข้อมูลผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูล



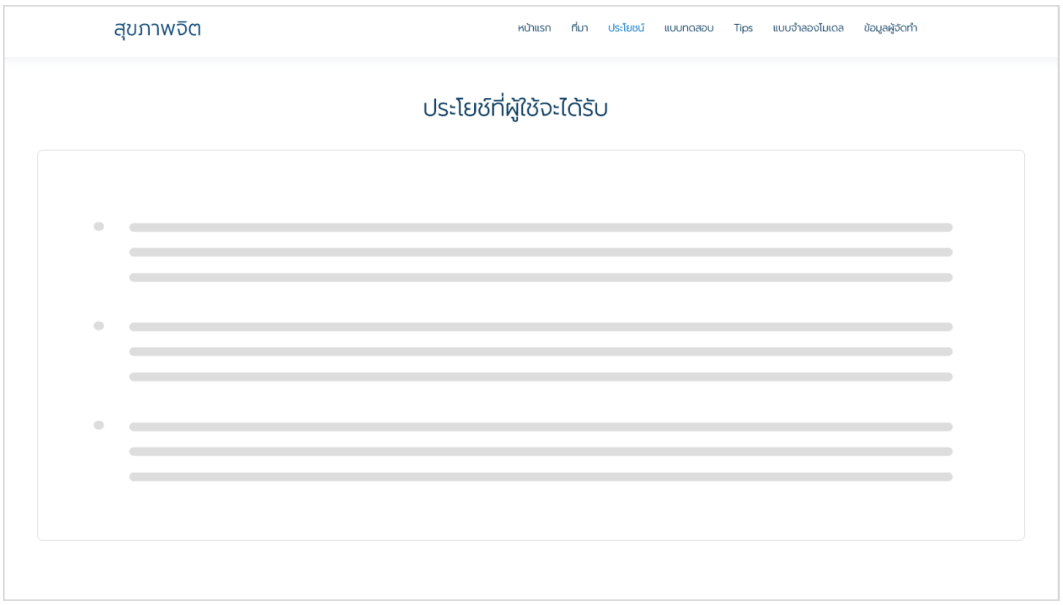
ภาพที่ 3.38 แสดงโครงร่างการออกแบบหน้าแสดงผลข้อมูลผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูล

6) หน้าแสดงผลข้อมูลที่มาของการวิเคราะห์ข้อมูลปัญหาสุขภาพจิตในการทำงาน



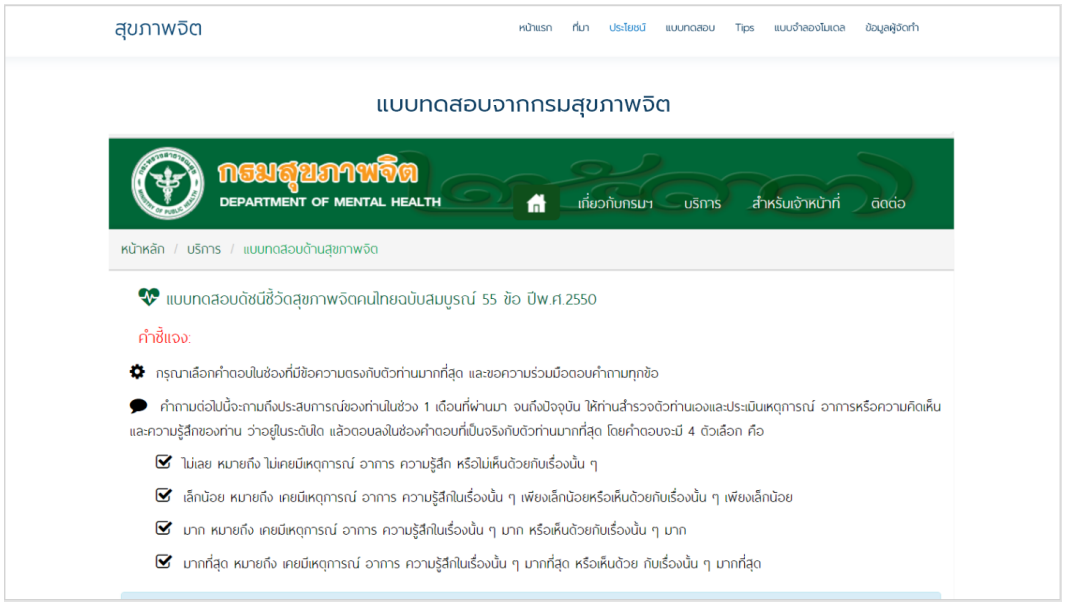
ภาพที่ 3.39 แสดงโครงร่างการออกแบบหน้าแสดงผลที่มาของการวิเคราะห์ข้อมูล

7) หน้าแสดงผลข้อมูลประโยชน์ที่ผู้ใช้จะได้รับของการวิเคราะห์ข้อมูลปัญหาสุขภาพจิตในการทำงาน



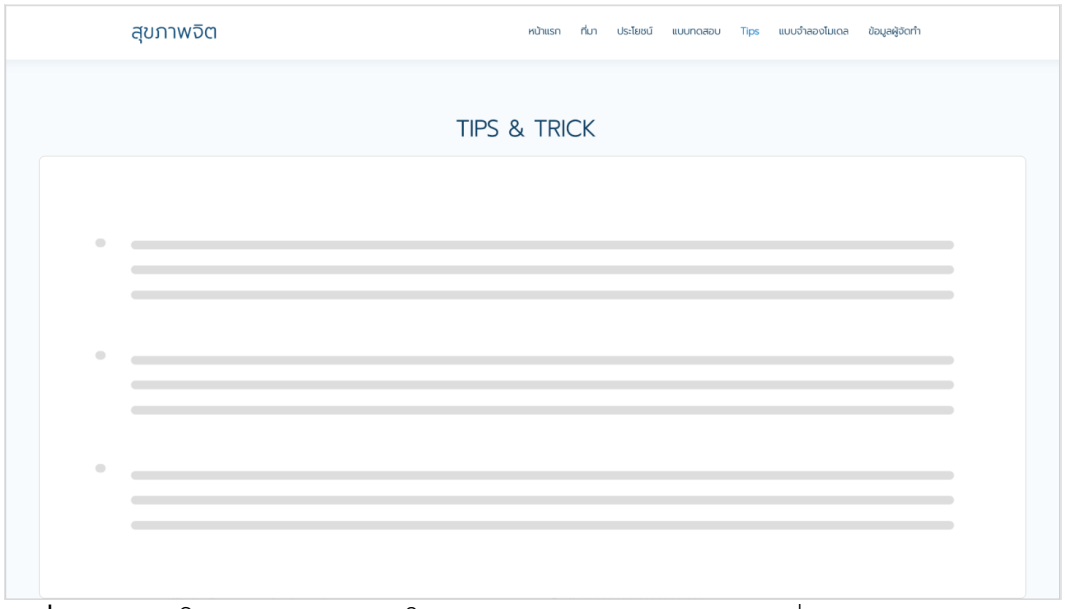
ภาพที่ 3.40 แสดงโครงร่างการออกแบบหน้าแสดงผลประโยชน์ที่ผู้ใช้จะได้รับ

8) หน้าแสดงผลแบบสอบถามดัชนีชี้วัดสุขภาพจิตในการทำงาน จากเว็บไซต์ของกรมสุขภาพจิต www.dmh.go.th/



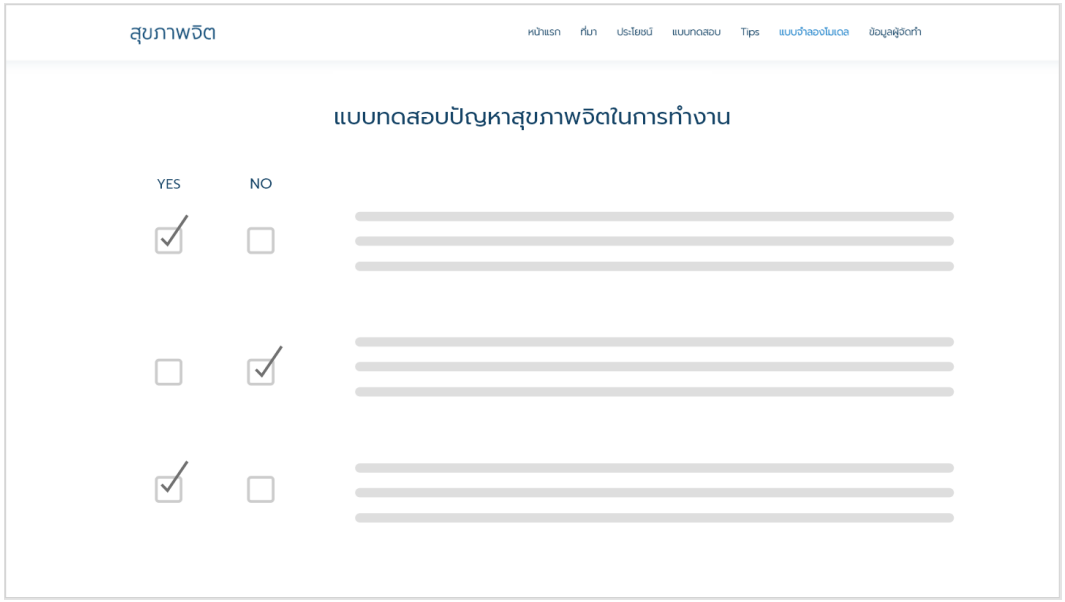
ภาพที่ 3.41 แสดงโครงร่างการออกแบบหน้าแบบสอบถามดัชนีชี้วัดสุขภาพจิตในการทำงาน

9) หน้าแสดงผลในส่วนของข้อมูลปัจจัยเสี่ยงและปัจจัยป้องกันปัญหาสุขภาพจิต



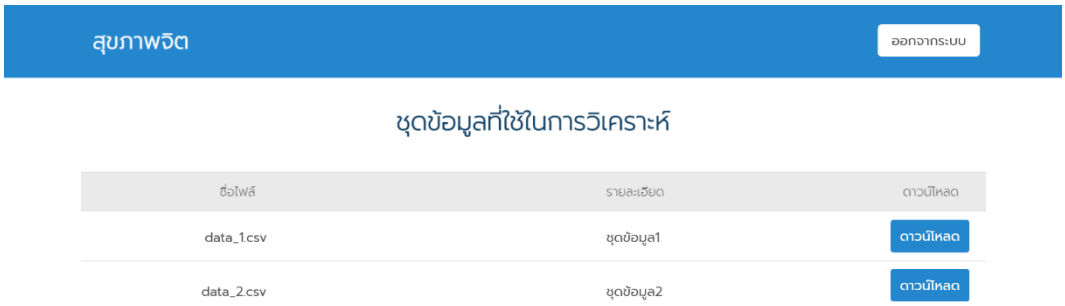
ภาพที่ 3.42 แสดงโครงร่างการออกแบบในส่วนของหน้าแสดงผลข้อมูลปัจจัยเสี่ยงและปัจจัยป้องกันปัญหาสุขภาพจิต

10) หน้าแสดงผลในส่วนของแบบทดสอบปัญหาสุขภาพจิตในการทำงาน



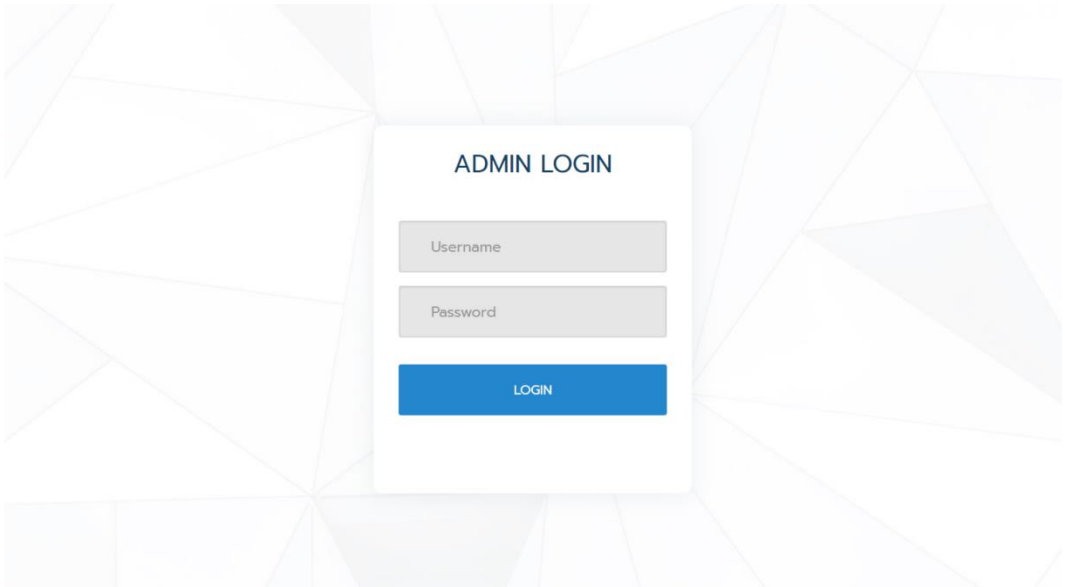
ภาพที่ 3.43 แสดงโครงร่างการออกแบบในส่วนของหน้าแสดงผลแบบทดสอบปัญหาสุขภาพจิตในการทำงาน

11) หน้าแสดงผลในส่วนของ Data Set เป็นหน้าแสดงชุดข้อมูลที่ใช้ในการวิเคราะห์ข้อมูล



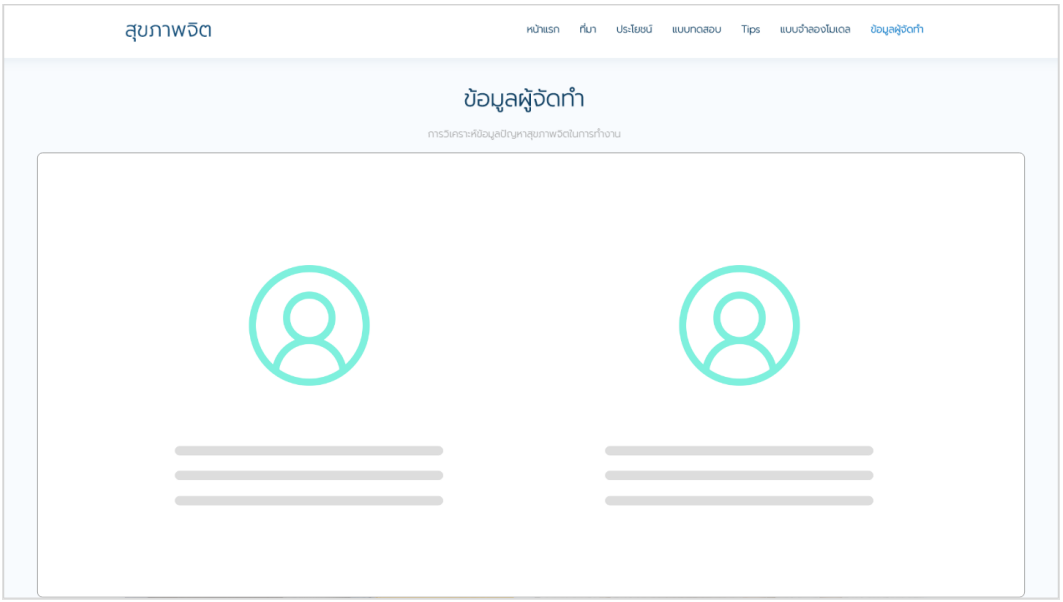
ภาพที่ 3.44 แสดงโครงร่างการออกแบบในส่วนของ Data Set

12) หน้าแสดงผลในส่วนของการเข้าสู่ระบบเพื่อไปจัดการกับข้อมูลภายในเว็บไซต์



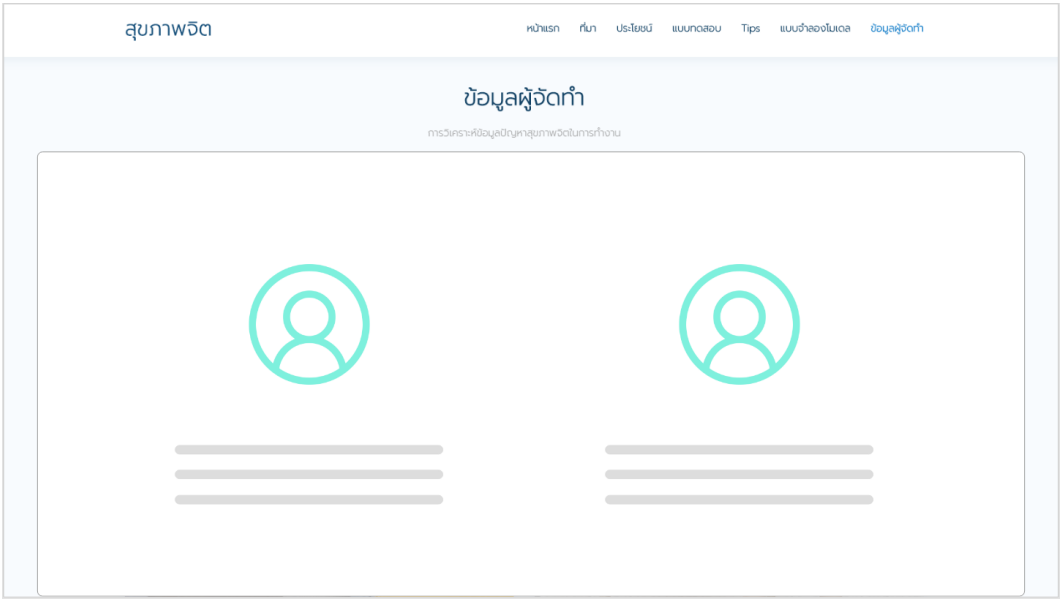
ภาพที่ 3.45 แสดงโครงร่างการออกแบบในส่วนของหน้าการเข้าสู่ระบบ

13) หน้าแสดงผลในส่วนของการจัดการกับข้อมูลภายในเว็บไซต์



ภาพที่ 3.46 แสดงโครงร่างการออกแบบในส่วนของการจัดการกับข้อมูลภายในเว็บไซต์

14) หน้าแสดงผลในส่วนของข้อมูลของผู้จัดทำ



ภาพที่ 3.47 แสดงโครงร่างการออกแบบในส่วนของหน้าแสดงผลข้อมูลของผู้จัดทำ

3.3 บทสรุป

จากวิธีการดำเนินงานโครงงานในข้างต้นทั้งหมดนี้ ผู้วิเคราะห์ข้อมูลได้แสดงวิธีในการจัดการกับข้อมูลปัญหาสุขภาพจิตในการทำงาน ประจำปี 2557-2559 ด้วยขั้นตอนกระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM อย่างละเอียด รวมถึงการสร้างโมเดล Decision Tree จากโปรแกรมที่ใช้เลือกทำเหมืองข้อมูลเพื่อใช้นำเสนอ คือ โปรแกรม Weka 3.8.4 ในการสร้างโมเดล Decision Tree และประเมินประสิทธิภาพของโมเดล ซึ่งมีค่าความถูกต้องเฉลี่ยในทุกโมเดลเท่ากับ 73.71% เป็นการทดสอบประสิทธิภาพของโมเดลด้วยวิธี Self Consistency Test พบว่ามีผลลัพธ์ตรงกันอยู่ในระดับค่อนข้างดี สามารถนำโมเดลไปใช้งานได้ และได้ผลลัพธ์ที่เหมือนกัน เมื่อเปรียบเทียบโมเดลกับโปรแกรม RapidMiner Studio 9.5.1 ทางคณะได้นำข้อมูลสารสนเทศมาทำการแสดงผลแบบ Visualization ในรูปแบบของแผนภาพและแผนภูมิชนิดต่าง ๆ โดยใช้โปรแกรม Power BI และออกแบบ Wireframe ของเว็บไซต์ที่จะเผยแพร่บน Web Browser ด้วยโปรแกรม Adobe XD